

Taking a hard line on conflicts

A clampdown on conflicts of interest at the US National Institutes of Health needn't stifle quality research at the agency — and it might indicate the shape of things to come elsewhere.

On 1 February, Elias Zerhouni, the director of the National Institutes of Health (NIH), announced new rules that are clearly intended to put an end to one of the more embarrassing episodes in the agency's recent history.

The rules put strict limits or outright bans on the outside activities of 18,000 NIH employees, most of them at the agency's main campus in Bethesda, Maryland. They were drawn up in response to a series of articles in the *Los Angeles Times* that spurred a congressional investigation into alleged abuses of the NIH's previous conflict-of-interest rules. Zerhouni says the rules have to be stringent, and that the agency must draw "a bright line" between unethical and ethical activities.

All the agency's employees will now be banned from doing paid or unpaid consulting for industry, healthcare providers, educational or non-profit institutes that receive NIH money, as well as for professional, trade or lobbying associations. Senior employees are banned from receiving many types of monetary gift, and from holding any stock in biotechnology, pharmaceutical and medical-device companies. Non-senior staff can only hold up to \$15,000 of such stock.

The new rules reverse those issued in 1995 by then NIH director Harold Varmus, which expressly permitted consulting activities as part of an effort to entice university scientists to join the agency. Some NIH staff are upset by the new restrictions, and many view them as an overreaction to abuses by a small number of scientists. Scientific societies are also expressing legitimate concerns about what the rules will mean for NIH scientists who want to serve on editorial boards, or take part in their professional development activities.

If Varmus's rules allowed too much outside commercial activity,

Zerhouni's seem to allow too little. Clearly, his back was up against the wall. The conflict-of-interest scandal first broke in December 2003, and he proposed a softer response. But as 2004 wore on, embarrassing revelations piled up, and it became clear that about 100 scientists had failed to inform the NIH of their activities, and a few were engaged in more egregious conduct. Publicity over the latter forced Zerhouni to take what he admits is "drastic action".

The restrictions will make it tougher for the NIH to recruit certain types of researcher. This may be especially true given other knocks to morale in recent years, such as the security measures that give the NIH the feel of a 'gated community' in the aftermath of the terrorist attacks of 11 September 2001. But a great many loyal NIH researchers are there not for the money, but for the chance to work at the largest, most diverse and best-equipped biomedical research centre in the world. It would be foolhardy to regard restrictions on outside commercial activity as the death knell for the agency's ability to attract top talent.

Indeed, the clampdown at the NIH could well foreshadow similar action outside the agency. Momentum is building for an examination of conflicts of interest on the campuses of research universities and academic medical centres around the United States. Last week, Zerhouni himself raised the issue: "In addition to preserving the trust of the public in a federal agency at the NIH, we need to preserve the trust of the public in the research enterprise in general." He added that there is a need to encourage a re-evaluation of the conflict-of-interest policies at academic campuses. With the nation's attention now drawn to the issue, scientists and institutions everywhere should be sure that their own houses are fully in order. ■

Not so fast

Anyone thinking of collaborations with emerging biomedical powers should test the ethical waters before jumping in.

A UK government 'Global Watch Mission' returned from a two-week trip to China, Singapore and South Korea in September with a glowing report on the state of stem-cell research there. A press release announcing the report, released last month, noted that these countries were "bringing their ethical standards more in line with the UK model" and suggested that "the quality and international standing of UK-based stem cell organisations would provide an ideal opportunity for UK–Far East collaboration".

All set then. But is this just wishful thinking?

Each of these countries recently put national regulatory frameworks in place, with South Korea's coming into effect just last month (see *Nature* **433**, 186; 2005). But are they observed and enforced?

Members of the UK party visited some top-class laboratories in China (www.globalwatchonline.com/missions/tmsmrep.aspx). But can their observations be generalized to the whole country? After noting that in China there is "much less resistance than would be met in the West to pursuing experimental therapies into clinical practice", the summary follows: "This is not to imply that clinical research is either unethical or unregulated."

Nor should it be taken as a sign that all is well. There are stories of

Chinese trials that fail to get informed consent and take advantage of desperate patients. China has regulations, but how well they are observed across the country is another question — indeed, many of those trials are said to involve local government officials. We have only to remember the blood-collection scandals in which uncounted numbers of Chinese were infected with HIV to know what local governments can do without anyone noticing — until it all blows up.

Anyone considering collaborations in a country with a questionable record on ethical regulation should snoop around first. The UK mission in China encountered repeated obstacles to inspecting informed-consent forms. The South Korean researchers who carried out last year's famed cloning experiments have also consistently refused to show the types of forms they used (*Nature* **429**, 3; 2004). These are warning signs.

The new regulations may snap researchers into shape in the same way that China's political measures ensure conformity. This strictness would then, it is hoped, extend to international collaborations. But the odds are that some researchers will take advantage of lax enforcement in China or other countries. Some will get away with it. Others might sink their careers in a project gone awry. ■





Trial stopped
Third cancer case
halts gene therapy
for children
p561



Heat is on
Wobbles in past
climate leave
forecast unchanged
p562



Kyoto calling
City prepares gala
as climate protocol
comes into force
p562



Core values
Chinese extract
key ice samples
from Antarctica
p564

Main agencies hang on to funds in skimpy US science budget

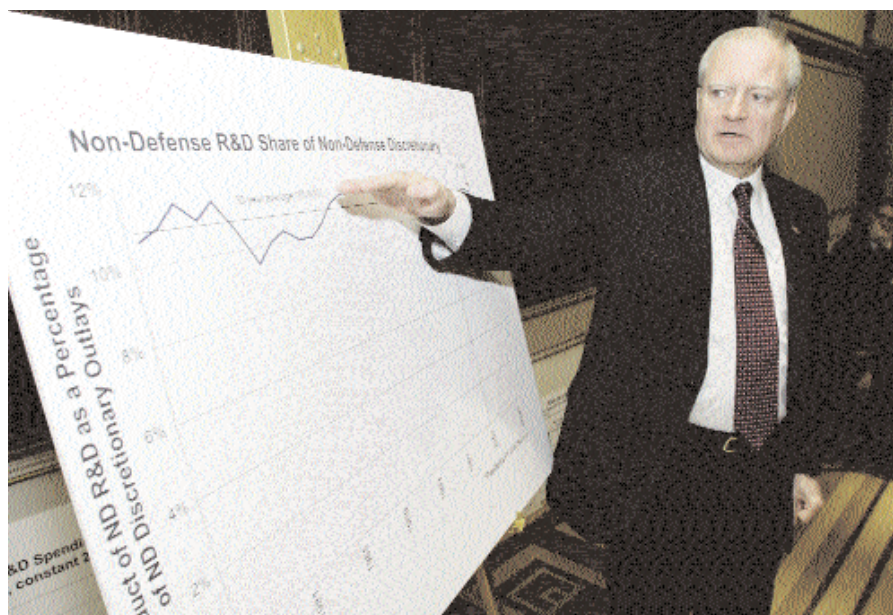
Geoff Brumfiel, Washington

US research funding looks set to take a hit this year as the government struggles to contain a swelling budget deficit while covering the rising cost of the Iraq war.

On 7 February, President George W. Bush laid out his US\$2.57-trillion plan for federal government spending in the 2006 fiscal year, which begins on 1 October. The proposed budget would cut "federal spending in science and technology" by 1.4% next year, to \$60.8 billion. This is the best measure for expenditure on innovative research and development, as defined by the US National Academies.

"The budget is not flat, but it's pretty close," said John Marburger, the president's science adviser, at a press conference in Washington. "There are some difficult cuts that will present challenges to science agencies and programmes." The budget has to be approved by Congress before it can become law.

The largest and most prestigious scientific agencies managed to hold their own in the budget proposal, however. Under the proposal, funding at the National Institutes of Health (NIH), for example, will rise very slightly to \$28.8 billion. But that modest increase is a far cry from the 15% hikes it enjoyed for several years until 2003 (see



Picture this: John Marburger, Bush's science adviser, presents the US president's budget for 2006.

'Disappointment in slow-down for biomedical funding', below). NASA, meanwhile, will see its budget grow by 2.4% to around \$16.5 billion. Much of that increase will be absorbed by preparations for future manned missions to the Moon and Mars (see 'Science squeezed by NASA focus on exploration', overleaf).

And the National Science Foundation (NSF), which supports most non-biomedical research at US universities and which some anticipated would face the first budget cut in its history, seems to have received a reprieve. Under Bush's request, its 2006 budget will grow by 2.4%, to \$5.6 billion. ►

Disappointment in slow-down for biomedical funding

The engine that drives US biomedical research may finally be running out of gas.

The National Institutes of Health (NIH) was held to a 0.7% budget increase in the 2006 proposal that President George W. Bush sent to Congress earlier this week. The biomedical agency will see its budget climb by \$196 million, to \$28.8 billion, if Congress approves the request.

That meagre increase does not come close to keeping pace with the 3.2% increase in biomedical costs that the government predicts for next year. So biomedical scientists funded by the NIH will increasingly feel the pinch.

Elias Zerhouni, the NIH director, tried to put a brave face on the situation at a press briefing on 7 February, pointing out that other domestic

programmes are being dramatically cut or even eliminated. "In relative terms," he said, "we're happy to have an increase. But it will now require tough choices to be made."

One of those tough choices includes the decision to fund 400 fewer postdocs than are being funded in 2005, although those who do get NIH support will be given slightly better salaries and more generous healthcare benefits. The NIH will manage to fund an extra 250 new and competing investigator-initiated grants, but as old grants expire the total number that the agency provides will fall by 400, to less than 39,000.

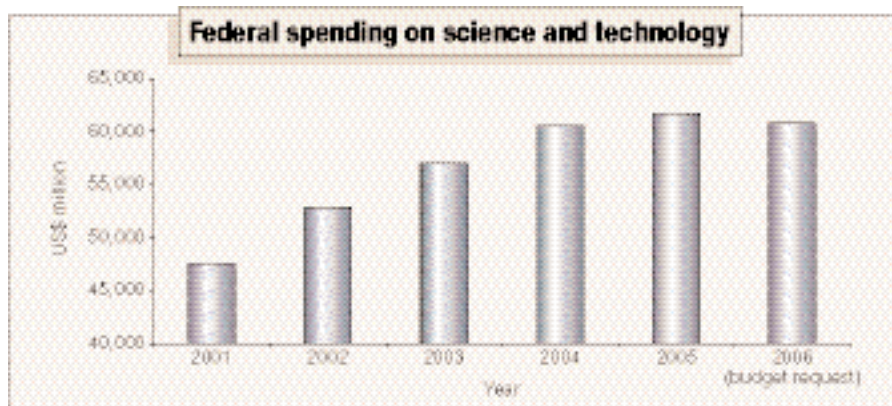
At the briefing, Bush's new health secretary, Michael Leavitt, defended the 2006 proposal,

pointing out that the NIH budget doubled over the five years from 1999 to 2003. "We have planted. It's now time for us to harvest the fruit," he said.

But research advocates expressed disappointment, saying that this budget will force the NIH to neglect the tremendous research opportunities generated by the doubling. David Moore, a lobbyist at the Association of American Medical Colleges, says: "It's going to be increasingly competitive for scientists to get even top-quality research funded" by the NIH.

"That's a pretty sobering message — particularly for young people considering careers in science," Moore adds.

Meredith Wadman



However, NSF grantees can expect tough times next year, close observers of the agency say. For example, \$43 million of the agency's \$132-million increase will be needed to operate icebreakers that clear the way to the NSF's polar research stations — an undertaking that until this year was paid for by the Coast Guard. This "isn't new research, it's a service that you need to have", says one critic of the administration. Another \$76 million of the increase is for major facilities, including EarthScope — a global network of seismic and other geophysical instruments — and IceCube, a neutrino observatory planned for the Antarctic.

In order to find money for additional research grants, the NSF intends to make deep cuts in programmes that support science education in US schools. The proposal reduces education funding by 12.4% to \$737 million. "Put the cuts together over the past two years, and the education budget is down by about 22%. They are clearly reducing the NSF's role" in school education, says Joel

Widder, a former agency official who works for a lobbying firm, Lewis-Burke Associates, based in Washington DC.

But in a budget that is fairly flat for science and technology overall, the modest gains at the NSF, NIH and NASA are offset by cuts elsewhere. The Department of Energy's office of science, for example, which funds most US physics facilities and research, will be cut by 4% to \$3.5 billion. Energy department laboratories will probably feel the brunt of this in their operating budgets. "We've known for some time that this would be a grim budget and now it's confirmed," says Tom Ludlam, a physicist at Brookhaven National Laboratory in New York state.

The Department of Defense, another major supporter of the physical sciences, is facing proposed cuts in its support of basic and applied research of 14% to \$5.5 billion. "This is not a happy situation," reflects Michael Lubell, head of public affairs for the American Physical Society.

Additional reporting by Emma Marris and Jessica Ebert.

Science squeezed by NASA focus on exploration

President George W. Bush's Moon-Mars programme marches onwards in his 2006 budget request for NASA: the proposal includes sharp funding increases for a new Crew Exploration Vehicle and robotic lunar missions to pave the way for future human travel in space.

But Bush's request on behalf of the agency for \$16.45 billion is just 2.4% more than its 2005 budget — half the level of increase that the White House promised a year ago. And that means something at NASA has to give.

One prominent casualty is the Jupiter Icy Moons Orbiter (JIMO), which was deemed too expensive and risky for the first use of nuclear propulsion (see *Nature* **433**, 342; 2005). NASA will now mount a less ambitious nuclear mission, and defer the Jupiter probe indefinitely. "We'd like to be able to do JIMO one day," says Craig Steidle, who heads NASA's Office of Exploration Systems.

The agency's newly merged Earth science and space science offices will see their budget drop by 1% to just under \$5.5 billion in 2006, if

Congress accepts the proposal. The increase of 2% to \$1.9 billion for Solar System exploration and increases for the robotic Mars and Moon missions (\$858 million, up 17% from last year) mean that other projects will be squeezed. For example, a New Frontiers mission to the outer Solar System is likely to slip, as is the launch of the Kepler probe, which will search for planets around other stars and was planned to launch in October 2007. Spending on advanced technologies such as aerocapture and solar electric propulsion will also be deferred.

Astronomy missions feeling the pinch include the Space Interferometry Mission, whose launch will be put back two years to 2012. A Joint Dark Energy Mission with the Department of Energy misses another chance for a new budget start. And although the James Webb Space Telescope remains on track for a 2011 launch, NASA is — for now — sticking with its decision to drop the Hubble telescope in the ocean when its batteries fail in 2008.

Tony Reichhardt

US budget in brief

Bioterror still top priority...

Efforts to counter bioterrorism remain a priority in this year's proposal, even as budgets tighten for other health programmes.

John Marburger, the president's science adviser, said that more than \$700 million would be spent over several years rebuilding biodefence research facilities at Fort Detrick, Maryland.

At the Centers for Disease Control and Prevention — which is taking an overall cut of 6% — funding for bioterrorism preparations will be boosted by \$56 million. The Food and Drug Administration will spend an extra \$30 million on preventing food contamination with bioterror agents. And the National Institutes of Health will inject \$97 million into developing counter-measures to nuclear and chemical threats.

...but no new nukes

The Department of Energy will ask for just \$4 million to fund research into an earth-penetrating nuclear weapon — a retreat from last year, when Congress rejected its request for \$28 million for the project.

The change suggests that the administration is giving some ground to critics of the 'bunker-buster' device. The money would be used only for "analysis" of the new weapon, not for its development, say department officials.

Rock solid

Extra money for earthquake monitoring and for the troubled Landsat Earth-viewing satellite are among the few upturns in an otherwise flat \$933-million request for the US Geological Survey.

Appropriations shake-up

As Congress starts to review the budget request, it is contemplating major reform of the powerful committees that determine science-agency funding.

Capitol Hill is buzzing over a proposal by Jerry Lewis (Republican, California), chairman of the Appropriations Committee in the House of Representatives, to dissolve 3 of the 13 appropriations subcommittees in the House and the Senate that carve up the budget between them. Among the subcommittees that might be axed are those dealing with funding for NASA, the National Science Foundation and the Environmental Protection Agency.

If this happens, these agencies will come under the jurisdiction of other subcommittees — possibly those dealing with energy and water, where they would compete directly with the Department of Energy for funds. The Republican leadership in Congress is expected to decide whether to pursue the reform by the end of next week.



Small but important: researchers hope that changes to a gene vector will reduce risks to patients.

NIH open-access plans draw fire from both sides

Erika Check, Washington

The US National Institutes of Health (NIH) has unveiled its long-awaited plan for open access to research findings. Elias Zerhouni, the NIH's director, claimed at a public briefing on 3 February that the plan could "change the landscape" of biomedical publishing.

The policy requests that authors whose research was funded by the NIH submit copies of their papers to the agency's National Library of Medicine after they are accepted for publication. The papers will then be placed in an online archive. Authors can decide when their papers are made available to the public, but the NIH would like this to happen as soon as possible, and in any case within 12 months of publication.

Scientists pushing for open access have praised the policy, which comes into effect on 2 May. "This is a significant and positive step and I'm glad we have the policy written down," says Harold Varmus, president of the Memorial Sloan-Kettering Cancer Center in New York.

But both sides of the debate have voiced criticism. Advocates of full open access are unhappy that the policy is voluntary and does not require access within six months of publication — a deadline that Zerhouni had proposed in a draft version (*Science* 306, 1895; 2004).

"This is a retreat from the earlier version of the policy, and the retreat is unjustified and regrettable," says Peter Suber, director of the Open Access Project at Public Knowledge, a non-profit advocacy group in Washington DC.

Suber and other critics say it would put researchers in the difficult position of having to negotiate between the NIH, which wants work available as soon as possible, and journals, which may want researchers to wait.

Publishers and societies that draw income from publishing criticize the NIH's plan to archive papers on its own site instead of directing the public to journal websites. NIH officials estimate that it will cost between \$2 million and \$4 million a year to run.

"The NIH is proposing to create a new publishing enterprise, and it's going to have to spend a lot of money to do that," says Marc Brodsky, chief executive of the American Institute of Physics.

Zerhouni said the announcement, expected on 11 January, had been delayed at the request of the health department and the White House.

Gene therapy put on hold as third child develops cancer

Erika Check, Washington

Scientists have halted clinical trials of gene therapy to treat a rare immune disorder — less than a year after the trials were relaunched following an earlier stoppage.

The trials use gene therapy to treat different forms of severe combined immunodeficiency disease (SCID). The first trial to be stopped was halted in October 2002, and other trials were halted three months later, after two children in the trials developed cancer. But authorities allowed them to resume during the past year because the treatment had cured many children who lack reliable alternative treatments.

Researchers have now halted the trials again, after a third patient was found to have developed cancer. The suspension is a significant setback for the nascent field of gene therapy, because SCID treatment has been its most promising application to date.

The child with cancer was a patient of Alain Fischer of the Necker Hospital in Paris. He has been using gene therapy to treat the X-linked form of SCID, which is otherwise only treatable with bone-marrow transplant and is still often fatal. Fischer's trial restarted last May, and his team has treated one child since then.

But on 24 January, the French medical

regulatory authority AFSSPS announced that a child who was treated by Fischer in April 2002 now has cancer.

As a result, Fischer's trial and similar ones in the United States have been halted again. The agency also said that one of the original two patients who had been diagnosed with cancer — both of whom were in Fischer's trial — died last October.

Fischer is now investigating why the third child, who was treated at a later age than the previous two children, developed cancer. The child's cells did not seem to have the same genetic glitch that caused the first two cancers, he says, but he cautions that the analysis is still under way.

Fischer adds that he still believes in gene therapy as a treatment for X-linked SCID, because 15 children treated in this way are still alive, and 14 are doing well four years later. But his group will not treat any more children using its current gene-therapy system, he says. He adds that he plans to change a key step in the treatment by changing the vector — the modified virus that delivers the therapeutic gene to the patients.

"The efficacy is there, but we have to improve on the safety," Fischer says, adding that this is "not an uncommon situation" in medical research.

Past climate comes into focus but warm forecast stays put

Quirin Schiermeier, Munich

Fluctuations in global temperature during the past millennium may have been larger and more frequent than previously thought, says a fresh analysis of the climate record.

The analysis is likely to reignite a long-standing controversy over the cause and extent of natural climate variability, scientists say, although the unprecedented nature of global warming since the mid-1980s remains unquestioned. The study was conducted by Anders Moberg of Stockholm University, Sweden, and his team (see page 613, and News and Views on page 587).

According to an earlier study, which produced the widely cited 'hockey stick' graph (see below), average Northern Hemisphere temperatures during the past millennium were relatively stable until the late nineteenth century, when they began to increase sharply¹. In 2001, this assessment was used to underpin the most recent report of the Intergovernmental Panel on Climate Change (IPCC) — the scientific branch of the United Nations Framework Convention on Climate Change.

But the Moberg study, which is published just as the Kyoto Protocol comes into effect (see 'Kyoto decks itself for celebration'), suggests that notable climate changes have occurred throughout the recent past. If such natural fluctuations continue in the future, they may "amplify or attenuate anthropogenic climate change significantly", the authors conclude.

Moberg's group used a combination of different 'proxies' to reconstruct decadal and centennial temperature changes. Proxies are climate indicators such as tree rings, pollens and boreholes, and the researchers used each one at the timescale that it records most accurately: tree rings are used for reflecting annual variations, for example, and sediments for longer-term changes. The researchers then used 'wavelet analysis' to combine the timescales in the optimum manner.

"At timescales longer than 80 years, temperature variability seems to have been considerably larger than previously thought," says Moberg.

Previously, different scientists had arrived

at different curves for temperature variability over past centuries, depending on the data or models they used². The possibility that they generally underestimated natural climate fluctuations has been one of the main arguments that sceptics use to reject the notion that human activity is responsible for current warming.

This argument has hardly any support in the climate community, however. Many researchers do agree that historic climate



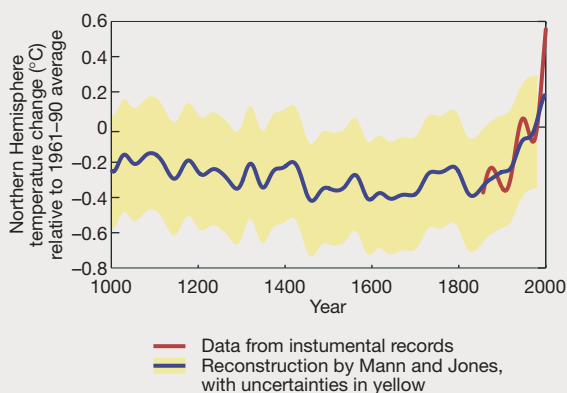
changes may have been underestimated. But the exceptionally strong warming trend since the mid-1980s cannot be explained by natural variability alone, they maintain. "Moberg's reconstruction will help to put the record straight in one of the most contested issues in palaeoclimatology," says Hans von Storch, a climate modeller at the GKSS research centre in Geesthacht, Germany. "But it does not weaken in any way the hypothesis that recent observed warming is a result mainly of human activity."

Moving on

"We need to understand the past, but some people become fixated," says Phil Jones, a climate researcher at the University of East Anglia in Norwich, UK. "For projecting the rate of climate change in the twenty-first century, it is somewhat irrelevant what happened in medieval times. What really matters is what happened in the twentieth century — and we can expect from that a much warmer climate."

In its 2001 report, the IPCC concluded that "the increase in temperature in the twen-

The 'hockey stick' graph



Kyoto decks itself for celebration as protocol comes into force

A colourful, international celebration in Kyoto on 16 February will mark the formal entry into force of the climate treaty that bears the city's name.

The Kyoto Protocol, negotiated in 1997, is already beset by scepticism over whether the signatory nations can meet the targets it sets for emissions of carbon dioxide and other greenhouse gases by 2012. And the agreement's effectiveness has been challenged by the refusal of the United States, by far the world's largest emitter of greenhouse gases, to participate.

But backers of the protocol hope that formal

implementation will spur efforts to reduce emissions. "We see 16 February as the start of a new era in international efforts to reduce climate change," says Joke Waller-Hunter, executive secretary for the United Nations Framework Convention on Climate Change, the 1992 treaty that the protocol supplements.

The ceremony will open with musical events and a talk detailing efforts taken to curb emissions. There will also be keynote speeches from Waller-Hunter and Wangari Maathai, Kenya's deputy environment minister and winner of last year's Nobel Peace Prize, as well as a live

message of support from Japan's prime minister, Junichiro Koizumi.

"The night will give high-level reconfirmations of member countries' commitments," says Takashi Ohmura, an official at Japan's environment ministry. "The protocol will pick up momentum."

It will need to. The latest data on greenhouse-gas emissions from the industrialized countries that have ratified the agreement suggest that many are falling short of their targets. In Japan, for example, two panels at the Ministry of Economy, Trade and Industry have found that 11 out of 30 industries surveyed, including steel,



Future summers are set to mirror 2003, when even the Netherlands enjoyed unusual heat.

tieth century is likely to have been the largest in any century during the past 1,000 years.”

Moberg’s reconstruction is consistent with this assessment. But, says van Storch, the hockey-stick curve, prominently featured in the IPCC’s summary for policy-makers, has become such a powerful icon that any correction of it will affect the credibility of the IPCC’s work. It could give climate sceptics a boost, despite the fact that human-driven global warming is not in doubt (see ‘UK climate meeting deems risks ‘serious’). The IPCC is likely to raise the issue in May in Beijing at a closed meeting of its working group on the physical basis of climate change.

The hockey-stick reconstruction was derived in 1998 by Michael Mann, a climate researcher now at the University of Virginia in Charlottesville. A small group of critics, including Stephen McIntyre, a Toronto-based mineral-exploitation consultant, has since attempted to prove that the graph is

based on insufficient data and on flawed statistics³. Although McIntyre’s work is controversial, a recent reanalysis by von Storch partly supports his view². And, in hindsight, many climate researchers believe that it was premature of the IPCC to give the visually suggestive curve so much prominence.

“Mann is a pioneer, whose 1998 study was then the best reconstruction that had ever been done,” says Stefan Rahmstorf, a climate researcher at the Potsdam Institute of Climate Impact Research in Germany. But, he adds, the controversy it generates is now out of proportion to its scientific significance.

Rahmstorf adds that even if the hockey-stick curve were to be completely wrong — and even if all model simulations of the past millennium were fundamentally wrong — it would hardly touch ideas about the cause of observed climate change in the twentieth century. Proxy-based reconstruction of past temperatures are important for validating the models that researchers use to predict the future climate. But, he says, “the cause of any particular climate change must be investigated separately. It would be naive to conclude that the observed twentieth-century warming must have a natural cause just because previous warming events have had one.”

Meanwhile, Mann concedes that it is plausible that past temperature variations may have been larger than thought — although he insists that Moberg’s reconstruction is not free of methodological and statistical problems. He says the issue deserves further investigation and must not be overshadowed by political issues.

“The contrarians would have us believe that the entire argument of anthropogenic climate change rests on our hockey-stick construction,” he says. “But in fact some of the most compelling evidence has absolutely nothing to do with it, and has been around much longer than our curve.” ■

1. Mann, M. E., Bradley, R. S. & Hughes, M. K. *Nature* **392**, 779–787 (1998).
2. Von Storch, H. et al. *Science* **306**, 679–682 (2004).
3. McIntyre, S. & McKittrick, R. *Geophys. Res. Lett.* doi:10.1029/2004GL012750 (in the press); pre-publication version at www.climate2003.com/pdfs/2004GL012750.pdf

power and electronics, would have difficulty meeting the protocol’s targets.

Eileen Claussen, president of the Pew Center on Global Climate Change in Arlington, Virginia, says that the jury is still out on the success of the agreement. The parties to the treaty “have negotiated extremely difficult targets”, she says, adding that the protocol’s implementation could call attention to this, and motivate stronger actions to reduce emissions.

Attention is now turning to what will happen after the target deadline in 2012 — but there is little sign of a rapprochement between those



nations that have ratified the agreement and those that have not. Ominously, the United States was not invited to send diplomatic representation to next week’s celebrations. “I don’t think they would have gone anyway,” says Claussen.

David Cyranoski, Tokyo

UK climate meeting deems risks ‘serious’

The risks attached to climate change are “more serious” than was supposed just four years ago, a meeting of top-level climate researchers has told the British government. The scientists add that “major investment is needed now” to mitigate the threat.

The meeting, held at the new headquarters of the UK Met Office in Exeter over 1–3 February, delivered a comprehensive summary of current understanding about global warming to its sponsor, the Department for Environment, Food and Rural Affairs.

Tony Blair, the British prime minister, is expected to use the advice to promote global action on climate change. He will take advantage of Britain’s presidency of the European Union, which will begin later this year, and its hosting of July’s meeting in Scotland of the G8, the group of eight leading industrialized nations.

But at the Exeter meeting, 200 climate specialists from some 30 countries found scant agreement regarding one government request: a tighter definition of what level of greenhouse gases in the atmosphere would constitute ‘dangerous’ human interference. Under the 1992 United Nations Framework Convention on Climate Change, the signatory nations — including the United States — agreed to act to prevent such a danger, but no one has defined what the word means in the context of the treaty.

Speakers at the meeting opined that levels of ‘danger’ were too variable between different regions for agreement. “There cannot be a definition of danger acceptable to everyone involved,” said Richard Tol, an economist from the University of Hamburg in Germany.

Governments, meanwhile, are unable to agree on an acceptable increase in global temperatures, much less an acceptable concentration of greenhouse gases. In several documents, the council of the European Union has said that a rise of 2 °C above pre-industrial levels should not be exceeded, although this point of view has not been widely accepted elsewhere.

At Exeter, Malte Meinshausen, a climate-policy expert from the Swiss Federal Institute of Technology in Zurich, said that models suggest it is “likely” that global temperatures will not rise by more than 2 °C if the level of greenhouse gases stabilizes at less than the equivalent of 400 parts per million of carbon dioxide — just 20 parts per million above today’s value.

The conference also identified research areas that deserve immediate attention, highlighting possible melting of the ice caps on Antarctica and Greenland, and the impact of carbon dioxide on ocean acidification.

Nicola Jones, Exeter

Swedish biobank data used to identify tsunami victims

Stockholm Swedish authorities have opened up a national biobank of blood samples in order to identify victims of the Asian tsunami.

The biobank, based at the Karolinska University Hospital in Stockholm, contains blood samples from babies born since 1974, and is used for medical research. But politicians approved a time-restricted law last month to allow police limited access to the biobank to match DNA samples with those from unidentified tsunami victims.

Concerns about the privacy of the biobank were raised in 2003, when it was opened to obtain a blood sample for a suspect in the murder of foreign minister Anna Lindh. This prompted some Swedes to demand that their samples in the registry be destroyed.

But in the tsunami identification case there have been no complaints, says Claes Guthenberg, head of the biobank laboratory: "We are seen as helping others in a disaster situation."

Sri Lanka turns to DNA test to pinpoint baby's parents

Tokyo Sri Lankan authorities are using DNA tests to establish the true parents of a baby washed up alive on a beach after the Asian tsunami in December. So far nine women have laid claim to Baby 81, who was the 81st admission to the hospital on 26 December. The results of the test are expected this week.

DNA tests are also being marshalled at public research institutes around the world for the more gruesome task of identifying victims of the tsunami. The Beijing Genomics Institute in China, for example, has already screened 100 samples from corpses from Thailand, and anticipates testing up to 4,000. Data on standard DNA markers that can be used to identify bodies and family relations will be deposited in a Thai data bank.

Medical agency blamed for upset over lab move

London UK politicians have criticized the Medical Research Council (MRC) for its handling of proposals to relocate one of the country's largest research institutes.

A report from a House of Commons select committee, due to be released this week, catalogues a deterioration in relations between the MRC and staff at the National Institute for Medical Research over the past two years. The MRC "signally failed to carry its troops with it" when discussing

China's hunt for climate clues reaches a peak

London One of the last unconquered corners of the world was reached last month by a team of a dozen Chinese scientists and explorers (see right).

The group travelled by tractor from the Chinese Antarctic base to the peak on the ice cap known as Dome A. At more than 4,000 metres above sea level, Dome A is the highest point on the continent's ice sheet. One of the team had to be helicoptered out when he developed suspected altitude sickness.

The scientists established a temporary station to extract ice-core samples, which provide a record of Antarctic climate. Dome A is important because it may yield cores older than the 900,000-year-old cores obtained by the European EPICA consortium. The EPICA core was recently extracted from a slightly less inhospitable site on the plateau called Dome C.

Astronomers also believe that Dome A has huge potential as a site for submillimetre, millimetre and infrared astronomy because of its dry conditions and clear skies. Plans to install a station at the site are under consideration in Australia and other countries.



YANG LEI/PHOTOCOME

proposals to move the institute from its present site at Mill Hill, on the outskirts of London, to a central site close to one of the capital's universities, concludes the report.

But the committee stopped short of endorsing allegations, made by some institute staff, that MRC chief executive Colin Blakemore had attempted to coerce



Colin Blakemore: wants to move Mill Hill lab.

members of a task force established to consider the proposed move (see *Nature* 432, 662; 2004). The council is due to make a decision on the proposals at a meeting scheduled for this week.

Washington state draws on tobacco cash for research

San Diego The newly elected governor of Washington state has proposed a \$350-million Life Sciences Discovery Fund to boost biomedical research. The move follows a trend of several US states hoping to ride the biotech wave.

Washington governor Christine Gregoire last week introduced measures in the state legislature to fund research to the tune of \$35 million per year for a decade, beginning

in 2008. The money is being made available from Washington's allotment of a national tobacco lawsuit settlement. Gregoire hopes that federal and private grants will take the total funding to more than \$1 billion.

Europe finds Viagra patent hard to swallow

Munich The drug company Pfizer has lost a seven-year battle for its 1998 European patent on the use of Viagra and related compounds in erectile dysfunction. The European Patent Office withdrew the patent on 3 February. But Pfizer's 1991 patent on the structure of Viagra's active ingredient, sildenafil citrate, is not affected by the ruling.

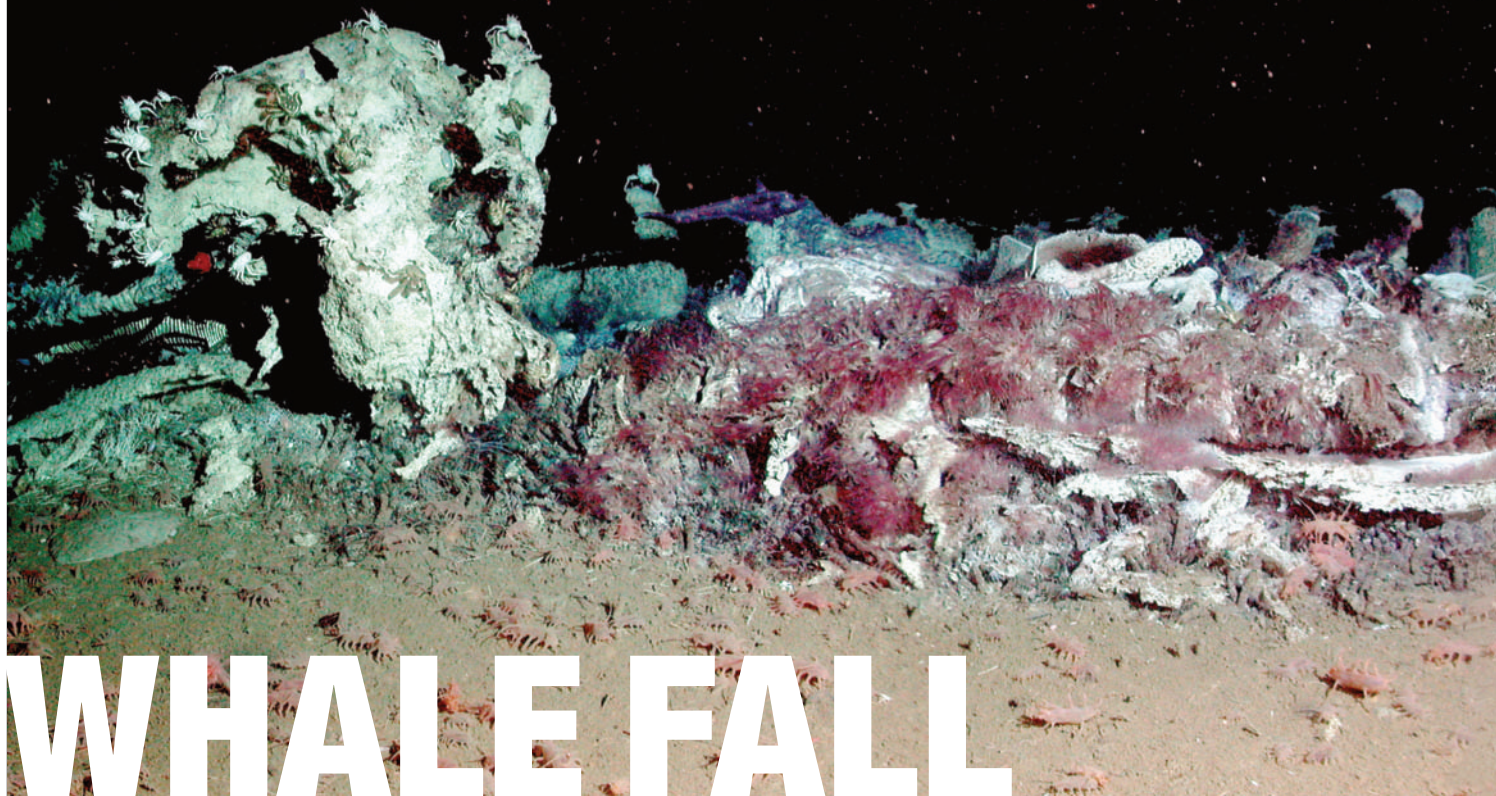
Sildenafil citrate belongs to a class of compound that inhibits a particular phosphodiesterase enzyme, leading to increased blood flow into the penis. Pfizer has sold more than one billion Viagra pills worldwide since it was introduced in 1998.

A consortium of competing drug companies argued that by the time Pfizer filed for the second Viagra patent in 1994, it was already widely known that such compounds improve erectile function. The patent office's decision allows competing companies to freely commercialize in Europe other drugs targeting the same enzyme. Two alternatives to Viagra — Bayer's Levitra and Eli Lilly's Cialis — have already been marketed pending the outcome of the patent battle.

Editorial note

Nature regrets that a scheduling oversight led to an inappropriate positioning of an advertisement from New England Biolabs (NEB) next to editorial coverage of the Asian tsunami, giving a misleading impression of insensitivity. Apologies to our readers and NEB.

MRC



WHALE FALL

The fatty bones of dead whales provide rich pickings for creatures on the sea floor. Amanda Haag meets the scientists who go to extreme and unpleasant lengths to study the unique ecosystems on these corpses.

In 1987, a manned submersible called *Alvin* was making a routine dive along the muddy plains of the deep sea when its pilot spotted what he thought was the fossilized remains of a dinosaur. Instead of an exotic underwater beast, it turned out to be the 21-metre-long skeleton of a blue whale. But atop this mass of bones the pilot did find something exotic: a carpet of creatures, including bacteria and worms, similar to those found on the flanks of underwater volcanoes.

The *Alvin* team had happened upon what have since become known as 'whale falls' — communities of creatures that thrive among the sulphur-laden ooze of decaying whales. Just as windfalls deliver a sudden bounty of ripened fruit, whale falls see the death of a whale bring a host of nutrients to the sea floor. The falls are few and far between, and difficult to track and study, but researchers are learning ever more — sometimes through extreme measures — about the new species to be found among the remains. Some 39 of the species discovered so far are thought to be especially suited or even unique to this environment.

In the barren depths of the open ocean, a fallen whale carcass is a veritable feast. Scavengers such as sleeper sharks, hagfish and squat lobsters can dine for months to many

years on the soft tissue of a single whale, depending on its size. As bits of whale tissue spread around the carcass, the enriched seafloor sediment provides nutrition for opportunistic worms and crustaceans. As anaerobic bacteria further the whale's decomposition, they create a sulphide-rich environment allowing sulphur-loving creatures to move in. Worms, clams and mussels, to name but a few, all take up residence, each getting their metabolic fix

from the chemical energy provided by the fat-rich marrow in whale bones¹.

Such complex communities have not been reported on the bones of other marine mammals, says Craig Smith, a whale-fall expert from the University of Hawaii at Manoa. This is probably because whale bones are so much larger and fat-rich, he says.

Scientists now estimate that a whale-fall community can survive for up to a century by sucking the fats and sulphides from these bones. The bacteria that make their home on whale falls are so good at degrading fat in cold waters that the biotechnology company Diversa in San Diego, California, is looking to see whether their enzymes

might prove useful in cold-water detergents.

For scientists such as Smith, studying whale falls presents some inherent challenges — not least of which is finding a body to examine. Many whales die when female whales or newborns, stressed by the process of birth, make their annual migration. In the Pacific, for example, there are often casualties on the long trek north to Alaska from birthing grounds along the Baja Peninsula of Mexico. Although whale carcasses tend to accumulate along these migratory paths, they fall at random locations and can be spaced very far apart.

Body hunt

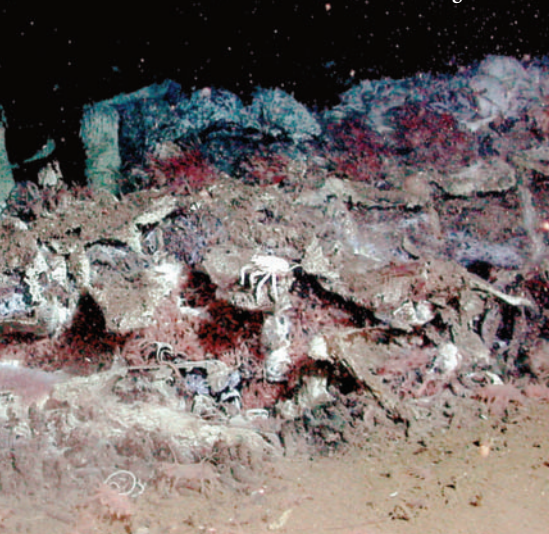
"They're hard to find because you can't just follow a particular geological feature on the sea floor and drive up to them like you can with hydrothermal vents," says Smith. In the early 1990s, the US Navy surveyed more than 300 square kilometres of the Pacific sea floor in search of a lost missile and found eight whale skeletons. The navy contacted Smith, but didn't take exact coordinates of where the skeletons were located. When Smith's team returned to the site, it could find only one of the eight.

So far, ten natural whale falls have been investigated by the dozen or so scientists who study them. And another 20 have been

"Whale bones are hanging in our cold room with cultures of bacteria stuck all over them."

— Thomas Dahlgren

Food for thought: the decomposing carcass of a whale can support a whole host of organisms.



Sinking feeling: intrigued by *Osedax*, a newly discovered genus of bone-eating worms (top and inset), and other odd creatures, scientists tow dead whales to the deeps for study.

sampled accidentally by trawling fishermen. In an attempt to increase these numbers, whale-fall researchers have resorted to sinking beached dead whales. But sinking a four-tonne juvenile grey whale, or a 35-tonne adult, is a major undertaking, says Smith.

When researchers get a call about a washed-up whale, they mobilize their teams and ships. By the time they reach the whale it is often bloated with gases from decomposition. To get it to sink, the scientists have to

weigh it down with up to 3,000 kilograms of scrap metal, from train wheels to anchor chains. It can take two days to get a whale from the shore to the sea floor — and all those involved agree that it is a highly unpleasant process. “We often throw away our clothes because you can’t get the smell out,” says Smith. “It’s one of the hazards of the job.”

Researchers have so far sunk about 20 whales this way. They return to these whale falls regularly to monitor the colonization of organisms on the remains. Smith’s group placed a time-lapse camera on one carcass, which took pictures more-or-less continuously for eight months. Others send submersibles to the sites, which can pick up rib bones and vertebrae and bring them to the surface with the communities still alive and mostly intact.

Thomas Dahlgren, a population geneticist at the Tjärnö Marine Biological Laboratory in Strömstad, Sweden, has been fortunate enough to sink a whale near his lab. The whale falls that have been studied off the coast of California tend to be 1,200–3,000 metres deep and at least half-a-day’s cruise away from the nearest lab; Dahlgren’s whale is 125 metres deep and just half an hour from his office. “We’re not stuck on a ship in the Pacific bobbing around with a limited time until the cruise is over,” says Dahlgren. He and his team can sample the whale fall as often as once a week.

Dahlgren’s lab also keeps whale bones with live specimens in a seawater tank. Unlike creatures found at hydrothermal vents, which can only survive at the high pressures found at the bottom of the deep sea, some whale-fall organisms seem to adapt quite well to life on the surface. “Whale bones are hanging in our cold room with cultures of these whale bacteria stuck all over them,” says Dahlgren.

Crunch time

Such research has unveiled a number of strange creatures. Perhaps the prize find so far is a newly described worm genus, *Osedax*² — Latin for ‘bone-devourer’ — which has a clever metabolic strategy. With no mouth, stomach or eyes, *Osedax* has evolved a root system to excavate the fat out of whale bones. The worms tunnel into the bones with their green, fleshy roots and turn them into “Swiss cheese”, says Smith. The worms then rely on bacteria

within their tissues to digest fats and oils from the bone marrow. Although the bacteria are similar to those found in oil slicks, this sort of microbe has never been found in a symbiotic partnership with another creature before.

So far, researchers have found five species of *Osedax* — four in the Pacific and, most recently, one in the Atlantic, implying that the worms have a worldwide distribution. Two of the species have a matriarchal society of sorts, in which all of the female members are about the length of an index finger, and the males are mere microscopic threads that live inside the females’ oviducts. A single female can hold up to 111 males. Another oddity is a whale-fall worm nicknamed ‘Pinky’, which at first evaded identification by its researchers, including Greg Rouse at the University of Adelaide in Australia. After much head-scratching, it became apparent that Pinky was just a polychaete worm — the class that includes ragworms and lugworms — albeit much larger than its shallow-water relatives. “Pinky is a giant,” says Rouse. “He’s more than one centimetre long, whereas his relatives are of the order of a couple of millimetres. Pinky’s also quite fat.” The worm measures a whopping one to two millimetres across.

Many mysteries about the whale-fall communities remain. For one thing, scientists are trying to find out how larvae from whale-fall organisms that are spawned into the water live long enough to find their way to another bone. “It is really clever. We don’t know how they do it, but they do it,” says Paul Tyler, a deep-sea ecologist at the Southampton Oceanography Centre, UK.

Scientists guess that the creatures in whale-fall communities probably date back some 35 million years. But in the past few centuries they have faced a serious threat. Smith estimates that whaling in the 1800s and 1900s reduced whale-fall habitat by up to 95%, potentially wiping out up to half of the species that were specialized to live on whale carcasses in some ocean basins³. Each time a whale was dragged aboard a ship, it not only depleted the live stocks, but also those of the dead falling to the sea floor. It’s not just the whales that needed saving, notes Smith. If hunting had continued apace it might have wiped out not only the great whales, but Pinky too.

■

Amanda Haag is a freelance science writer based in Boulder, Colorado.

1. Smith, C. R. & Baco, A. R. *Oceanogr. Mar. Biol. Annu. Rev.* **41**, 311–354 (2003).
2. Rouse, G. W., Goffredi, S. K. & Vrijenhoek, R. C. *Science* **305**, 668–671 (2004).
3. Smith, C. R. in *Whales, Whaling and Ocean Ecosystems* (ed. Estes, J.) (Univ. California Press, Berkeley, in the press).

MBARI

G. ROUSE

C. R. SMITH



Cold turkey, Vietnamese style

It was invented by a healer familiar with the horrors of opiate addiction, and refined by Vietnam's leading chemistry lab. Can this novel herbal cocktail ease withdrawal and reduce drug cravings? Peter Aldhous investigates.

Tung doesn't ask for much in life: he dreams of finding a job, getting married, and becoming, as he puts it, a "normal person". But the dark stains on his arms mark him out as a social pariah. Tung was once so desperate for his regular fix of opium that he would first smoke the drug, and then mix the residue with water to create an injectable paste that has discoloured his veins. Now 28, this is Tung's second time in rehab.

In the West, doctors might try to wean Tung off drugs by putting him on a substitute opiate such as methadone, and then gradually reducing the dose. But this is Vietnam, where the authorities take a tough line on what they view as a social evil. At the state-run drug-treatment clinic in Hoa Binh, some two hours' drive into the hills west of Hanoi, abrupt detoxification is the order of the day. Recovering addicts must first go 'cold turkey'

in concrete cells furnished with only a mattress. Many, including Tung, have been forcibly detained.

The conditions may be harsh, but the Hoa Binh clinic is at the centre of an experiment in drug rehabilitation that has sparked the curiosity of clinicians worldwide. Here, hundreds of addicts have been treated with a cocktail of herbal extracts called Heantos. It is claimed that it makes withdrawal tolerable and, in the longer term, reduces the craving for drugs.

From its crude beginnings as a thick brown syrup, Heantos has been refined into a series of subtly different formulations, given at different stages of detoxification and recovery. They are produced in a pilot plant at the Vietnamese Academy of Science and Technology's Institute of Chemistry in Hanoi. And within a year, a clinical trial in Germany may reveal whether the hopes that the

institute has invested in Heantos are justified.

In many countries, Heantos would have remained on the medical fringes. But in Vietnam, there is little of the tension that, in the West, divides the practice of herbalism from the high-tech world of pharmaceutical chemistry. So in the mid-1990s, when Tran Van Sung, now director of the Institute of Chemistry, was approached by a herbalist touting a concoction that could allegedly reform hardened drug addicts, his first instinct wasn't to dismiss the healer as a quack.

Refined approach

Instead, Sung engaged with the herbalist, Tran Khuong Dan, as a collaborator. "We had difficulties understanding one another in the beginning," Sung admits. Dan's focus on the overall blend of plant extracts in his creation initially jarred with Sung's desire to standardize and improve techniques for extracting the main compounds present. But slowly, their goals coalesced. Dan is today an associate of the chemistry institute, and is named as co-inventor, with Sung and his colleagues, on a German patent application.

The Heantos story began more than a decade before Dan's initial contact with Sung, when Dan was practising in Ho Chi Minh City. After seeing his brother and father fall foul of drugs, Dan travelled the country, investigating remedies used to treat addicts. Individually, none worked very well. But by the end of the decade, Dan had come up with a cocktail of herbal extracts that — judging from tests he conducted on himself, after deliberately getting hooked on various

HOANG DINH NAM/GETTY IMAGES



Social pariahs: Vietnamese opium addicts are shunned by society, but it is claimed that a cocktail of herbs has helped some in state-funded rehabilitation (below) to beat their problem.



opiates — provided relief from drug withdrawal. “It was dangerous, but there was no other choice,” says Dan.

Whether this tale has been embellished is hard to say. But flashy showmanship doesn't seem to be Dan's style. On our trip to Hoa Binh, he is quiet and detached; facts about his work on Heantos have to be coaxed from him through an interpreter.

Perhaps Dan has grown tired of telling the story, given the tortuous road that Heantos has travelled. The treatment first came to international attention in 1997, when the United Nations Development Programme (UNDP) decided to invest US\$400,000 in its scientific development. Since then, a similar

sum has come from the Norwegian government, with additional support from the United Nations Educational, Scientific and Cultural Organization.

By local standards, this is lavish funding. And it has caused some tensions within Vietnam, where there are competing approaches to drug detoxification. “Heantos has certainly been contentious,” says Ayo Wahlberg, a graduate student at the London School of Economics who is interested in the regulation of herbal medicines, and has made several trips to Vietnam. In the late 1990s, for instance, rumours circulated that some patients had died following adverse reactions. But subsequent safety evaluations by the Vietnamese authorities have given Heantos a clean bill of health.

In parallel, Sung and his colleagues have refined the techniques used to produce Heantos, replacing Dan's method of grinding the plants in water with solvent extraction and freeze-drying. They have also begun characterizing the chemical constituents of the resulting formulations. “We have several hundred pure compounds,” says Sung. But their identity, and that of the 13 plants from which Heantos is derived, is being kept under wraps until the German patent application is published.

Shortly after the UNDP got involved, there was a flurry of publicity when a group of foreign journalists interviewed Dan and met addicts who had been treated with Heantos. Some of the resulting articles painted Heantos as a miracle cure, citing positive results from Vietnamese trials — including a study of 110 veterans of what is known here as the American War, who had become addicted to the morphine used to ease the pain of their wounds.

Outside influence

But the Vietnamese clinical studies of Heantos have all failed to meet the gold standard of double blind, placebo-controlled trials. Clinical research is weak in Vietnam. So if the effectiveness of Heantos was to be assessed to internationally recognized standards, it seemed that this would require a trial outside Vietnam.

Initially, this was going to be carried out by a team led by Donald Jasinski at the Johns Hopkins School of Medicine in Baltimore, Maryland. But the planned trial's US proponents failed to secure funding. Eventually, the government of the state of North Rhine Westphalia in Germany, where attitudes to herbal medicine are more positive than in many other Western countries, came to the rescue. In 2001, it agreed to fund a trial to test the ability of Heantos to ease withdrawal symptoms.

The protocol, designed by psychiatrist Norbert Scherbaum of the University Hospital in Essen, calls for the treatment to be given

to 30 heroin addicts, while another 30 are given a placebo. But before the trial could go ahead, the German authorities demanded analysis of the Heantos formulations by the Institute of Plant Biochemistry in Halle.

Given that Heantos is such a complex mixture, satisfying the demands of regulatory authorities has been a persistent problem, says Wahlberg. Even a 2004 European Union (EU) directive, designed to lower the hurdles for certain herbal remedies, doesn't provide much assistance. It offers a simplified registration procedure, with few requirements for lengthy preclinical tests and clinical trials, for products that have “a well-established medicinal use” within the EU. Heantos is the product of Dan's invention and Sung's chemical expertise, so it doesn't qualify.

Scherbaum hopes to begin enrolling patients in the trial within the next few months, and to have results by the end of this year. Experts who have followed the progress of Heantos are taking a cautious view. “I've seen a lot of ‘magic bullets’ come and go,” says Gabrielle Welle-Strand of the Centre for Medication-Assisted Rehabilitation in Oslo, Norway, who was a member of an international team of addiction specialists that visited Vietnam in 1999 to witness the treatment's use.

Scherbaum's trial also isn't designed to test the claims that Heantos can reduce the underlying craving for drugs that prevents many reformed addicts from keeping clean in the long term. Here, there is no such thing as a miracle cure — as Tung can testify. After his first spell in rehab, aided by Heantos, he returned to his hometown of Mai Chau, in the mountains near the border with Laos. Unable to find work, Tung soon fell in with his old drug-taking friends, and was enslaved to opium once more.

But if Heantos is shown by Scherbaum's trial to have some therapeutic value, it would provide a boost to the prospect of a treatment that seems for the past few years to have become stuck in limbo. Even the Vietnamese authorities have yet to decide whether to approve Heantos for general use.

There's certainly a growing demand in this part of the world for treatments to aid drug rehabilitation. A few years back, Vietnamese drug addicts were few in number, and mostly rural users of opium. Today, there's a burgeoning heroin problem, both in the rapidly expanding cities and along the roads from Laos, used to transport the drug for illegal export.

This trade blights the region, and the wider world. Will Dan and Sung's efforts one day allow Vietnam to be known for exporting an antidote to addiction, rather than merely being a trade route for the criminals who feed it? The answer may lie in Essen. ■

Peter Aldous is *Nature's* chief news & features editor.

KATZ PICTURES

D. MARTINEZ/REUTERS/NEWS.COM

Prison talk

A few French scientists are bringing astronomy to captive audiences, such as the terminally ill and the incarcerated. Alison Abbott joined a group of convicted murderers to learn about gravity.

There is, it must be said, a certain irony in talking to the incarcerated about the stars and the infinity of space. And a particular poignancy in telling them about black holes — from which nothing can escape. Yet 50 or so prisoners, many of them lifers, at the maximum-security Muret Prison near Toulouse, France, listened with rapt attention to astrophysicist Didier Barret's 90-minute lecture on Einstein's theories of matter, space and time.

If Barret, who works at the Laboratory of Space Astrophysics (CESR) in Toulouse, was aware of these sensitivities, he did not show it. He marched slickly but relentlessly through the theory, to observations of the birth and collapse of stars, and on to his own front-line research on the celestial explosions known as γ -ray bursts. The prisoners didn't flinch — although occasionally one would walk out, this also being the day for family visits.

This was the last in a series of lectures at Muret — France's 'four-star prison', where inmates have their own rooms, and the warders encourage educational events. Barret, who organized the talks, believes in taking public outreach one step beyond the conventional. "There are some groups of people who physically can't come to a public lecture at the local science centre or wherever — so we take our public lectures to them," he says.

To pursue this philosophy Barret established an association called *Les Étoiles Brillent pour Tous* ('The stars shine for everyone') in May last year — with the Muret lectures as one of its first ventures.

The experience at Muret turned out to be less alarming than the lecturers had anticipated. Barret is a burly amateur soccer player, but confesses to being a little nervous when he first made his way through the seven security gates that divide each ring of the prison complex to reach the highest-security core. Here is located a round concrete church, where the lectures were to take place a few weeks later. Muret is a prison for serious sexual crimes and murder. It is not an easy place for intellectual curiosity to take root.

On his first visit, Barret remembers find-



ing the stage of the church a mess of broken furniture, cigarette ends and other rubbish. "I told the prisoners who were showing me around that it would not be possible to give a serious lecture in such surroundings," he says. They promised that they would clear it up in time for the first lecture — on astronomical dimensions — in June. And they kept their word. The stage was immaculate, and there was even a bottle of water, with a glass, for the speaker, Peter von Ballmoos, one of Barret's CESR colleagues.

Star pupils

Von Ballmoos had been warned by warders that prisoners often shout or jeer at guest speakers. "I had thought I would feel frightened and vulnerable on stage," he recalls. But instead he found them courteous and well-behaved, apparently won round by the seriousness of the subject.

The other four speakers in the series had similar experiences. Even the two female lecturers found themselves at ease despite the presence, within arm's reach, of an audience convicted of heinous crimes.

Discussions after the lectures, they said, were much like those after any 'normal' public lecture, although they tended to drift towards philosophical aspects of cosmology, such as the unimaginable scale of the Universe. Perhaps cosmology helps the prisoners gain a fresh perspective on their incarceration, muses von Ballmoos. "On such an immense scale the prisoners have essentially the same level of freedom as a non-prisoner," he says.

And whatever the topic of the lecture, the prisoners were always keen to discuss extreme events, such as the Big Bang and black holes — violent cosmological phe-



Didier Barret is taking science to inmates at Muret Prison (inset).

nomena in which matter gets crushed to infinite density under the pull of infinite gravity, and the laws of physics break down.

There were occasionally aggressive interjections that would not be expected in other public forums — for example, when one prisoner was irritated by a question from another and stormed out, yelling offensively; or when a technical problem with the microphone precipitated ill-tempered (and nerve-racking) heckling.

But the regular attendees tended to be well-behaved and deeply engaged. Talking after Barret's lecture, one prisoner, a former trade-union employee who had already served seven years, explained how physics became an immediate passion when he heard about CERN, the European particle-physics laboratory in Geneva, through his job. The prison lectures are magnificent, he says, and help him continue his self-education. A bit of a Renaissance man, he spends some of his prison time writing poetry, and the cosmos provides him with a rich source of metaphors.

Not all of the prisoners share this romantic outlook, but the lectures have been deemed such a success that the prison authorities are keen for a second series this year. And another prison in the area, the Maison d'Arrêde Seysses, has persuaded Barret to arrange a lecture series for it as well.

Les Étoiles Brillent pour Tous is there for the innocent as well as the guilty. Barret and his colleagues have given up their time to teach lower-level astronomy to terminally ill children who will never again see the outside of a hospital. "We feel an academic duty to bring our subject to every human being, whatever their disadvantages," says Barret. The stars, after all, shine for everyone. ■

Alison Abbott is *Nature's* senior European correspondent.

Nuclear nations should take the lead in disarming

A willingness to use 'overwhelming force' encourages other states to arm themselves.

Sir — We are glad to see that *Nature* is taking a strong line about the growing dangers of nuclear proliferation ("We have the technology" *Nature* 432, 432–437; 2004). When the Berlin Wall came down it was easy to become complacent: most people believed that the dangers of nuclear weapons had receded. That has proved overoptimistic. *Nature* is right to point out that the increasing number of states possessing, or on the point of possessing, nuclear weapons gives grave cause for alarm. We strongly support the view that scientists have a special responsibility to restrain these trends.

In his Commentary in the same issue, C. P. Robinson ("Revisiting the Baruch Plan" *Nature* 432, 441–442; 2004) takes the view that the nuclear non-proliferation treaty (NPT) is not the right tool with which to tackle this problem because of its "basic structural problems", which he says the world is not yet ready to address. He

proposes, instead, a gradual strengthening of regional alliances. This should be given careful consideration. But let us not forget that the original nuclear powers, when they signed the NPT, made an "unequivocal undertaking to accomplish the total elimination of their nuclear arsenal". We believe that the right approach is for the nuclear powers to abandon those policies that are incompatible with the letter and the spirit of the NPT.

For example, Robinson's claim that "The United States has been leading an international effort to reduce existing nuclear stockpiles, including its own" is not supported by talk of new 'bunker-busting' nuclear weapons. Congress has denied funds for 2005, but the administration has renewed its request for 2006. Furthermore, the administration has affirmed — in its "National Strategy to Combat Weapons of Mass Destruction", December 2002 — its willingness to use pre-emptive measures

against WMD-armed adversaries (a course contrary to Article 2(4) of the UN Charter) and to respond to an attack with overwhelming force, including use of WMD.

We believe that the failure of the official five nuclear states to abide by their NPT obligations encourages proliferation, by providing an excuse for other nations to acquire nuclear weapons. The United Kingdom, for example, should phase out its 'independent' nuclear weapons, such as Trident submarine-based systems.

Robinson claims that the United Nations "is not now an organization that can satisfactorily demonstrate security leadership". The United Nations must be brought up to date, but that can only happen with US support and US willingness to take a less self-centred and more global view than it is doing at present.

Robert A. Hinde

British Pugwash Group, 63a Great Russell Street, London WC1B 3BJ, UK

Military alliances offer no nuclear security

Sir — We welcome your Editorial ("Conscience call" *Nature* 432, 421; 2004), calling for active engagement by scientists in addressing the problems of nuclear-weapons proliferation. But the 'solution' proposed by C. P. Robinson of the Sandia National Laboratory in his Commentary article, "Revisiting the Baruch Plan" (*Nature* 432, 441–442; 2004), is an unlikely route to safety. He suggests more military alliances, each with a nuclear umbrella state providing 'security', forming a worldwide network of nuclear deterrence.

Would it not be more effective, and safer, to build alliances of nuclear-weapons-free countries? There are already several large and legally recognized nuclear-weapons-free zones (NWFZ). These include the areas covered by the Treaty of Tlatelolco in Latin America, of Rarotonga in the Pacific, of Bangkok in Southeast Asia and (yet to enter into force) of Pelindaba in Africa. Some states, including Mongolia and New Zealand, have individually declared themselves nuclear-weapons-free. Proposals have also been made for NWFZ in central Europe and the Middle East.

These treaties or decisions bind their adherents to non-nuclear policies. Some are recognized by the nuclear-weapons states, which agree not to use nuclear weapons against countries in the NWFZ.

If, as Robinson suggests, most or all states were linked to nuclear military

alliances, this would render them all liable to nuclear attack. NWFZs will tend to reduce the proliferation of nuclear weapons, and simultaneously make such countries safer by diminishing the overall likelihood of nuclear conflict.

In our view, the way forward is to treat nuclear weapons like other weapons of mass destruction (WMD), which have all involved scientists in their development. Chemical and biological weapons are forbidden by treaty. There exists a model Nuclear Weapons Convention, introduced into the United Nations in 1997 by Costa Rica. Its entry into force would end the hypocrisy of the nuclear-weapons states lecturing other countries for their nuclear ambitions. All WMD would be forbidden; there would be no need for WMD-dependent alliances; the entire globe would be a nuclear-weapons-free zone.

Dominique Lalanne*, **Peter Nicholls†**,
Joseph Rotblat‡

* *Accélérateur Linéaire, 91898 Orsay, France*

† *University of Essex, Colchester, Essex CO4 3SQ, UK*

‡ *8 Asmara Road, London NW2 3ST, UK*

Would you accept advice from a believer in Santa?

Sir — I was horrified to read the recent Editorial "Where theology matters" (*Nature* 432, 657; 2004) in the world's foremost science journal. Not only did the Editorial appear to support the position

that science and religion deal with different aspects of reality (which they do not: for example, either Jesus rose from the dead or he didn't — clearly a scientific question), but it also implied that religion has some privileged position in ethical debates.

This view is reflected in many public discussions of moral issues, where the obligatory priest or rabbi is wheeled out to comment on some topic, in spite of their utter lack of any qualification other than a belief in a paranormal entity that created the Universe and all it contains. Would you be prepared to accept fundamental advice from someone who insisted that Father Christmas was real?

The suggestion that religion has an intrinsic and predestined role in any ethical debate is indefensible, as a simple read of the ethics promulgated by the Old Testament (for example) will make abundantly clear. In most parts of the world, and certainly in the Western world, we no longer stone adulterers to death, the sins of fathers are not paid for by their sons, and masturbation is not viewed as a mortal sin.

It is also time for us to discard other atavisms, including pandering to religion and pretending that this out-dated, dogmatic endeavour is preordained to lead or advise us on any issue, ethical or otherwise. Dogma is not ethics.

D. J. Hosken

Centre for Ecology and Conservation, The University of Exeter in Cornwall, Tremough, Penryn, Cornwall TR10 9EZ, UK

Lessons from the past

Poverty and market forces combine to keep rural China unhealthy.

**Zigang Dong, Christina W. Hoven
and Allan Rosenfield**

The outbreak of severe acute respiratory syndrome (SARS) in China in 2003 and its subsequent spread to 30 other countries caused great alarm worldwide, and opened China to a new level of public scrutiny. After the outbreak, the Chinese government and the global community began to look at China's current health infrastructure and its leadership in more detail. China also faces a projected AIDS epidemic, which could infect up to 10% of the population by 2015. Sadly, as China has got richer, the inequalities of its health system have grown worse. And although the Chinese government and its Ministry of Health are now taking steps to rectify this, more emphasis must be placed on the health of the poor in rural areas, and new ways to fight greed and corruption are needed.

Despite these tremendous difficulties, we are optimistic that China can revitalize its collapsing public health system. We must not forget that between 1950 and 1970 China achieved much more in public health — with far fewer resources. Today, China is one of the world's largest economies, and thus has the capacity to face its public health challenges — if it decides to do so. The Chinese response to SARS shows that the government can respond swiftly and effectively to critical health needs, especially if persuaded to do so by external or internal public opinion.

The past: glorious beginning

After the formation of the People's Republic of China in 1949, the Chinese government created a highly efficient public health system using very limited resources, which was designed to be in harmony with Chinese health practices. The system involved practitioners of both traditional and western medicine. The national health policies of Chairman Mao Zedong resulted in several highly successful campaigns, which made China a role model for other developing countries to emulate¹.

The quality of health care in urban areas of China usually far surpassed that available in rural areas. To correct this inequity, from 1949 to 1980 China attempted to provide equitable distribution of adequate health care to all. A 'cooperative medical insurance system' was introduced into rural China to support three tiers of health care². At the first level, the government provided six months to one year of training for 'barefoot doctors'. These health care providers worked in village health centres,



Life's lottery: urban Chinese expect western-style health care, whereas the rural poor rely on cheaper clinics.

applying basic medical knowledge relevant to common diseases. At the second level were township hospitals and health centres. These facilities usually had 10–40 beds with qualified staff and a few doctors, who had trained for 3–5 years in medical schools. Only the most seriously ill patients were referred to the third tier, the state enterprise hospitals, which were supplied with the best equipment and staffed with doctors who had spent a minimum of five years at medical schools.

In addition, the government introduced systematic 'prevention first' health campaigns against specific diseases. As a result, public health in China improved dramatically. Many epidemic diseases, such as plague, cholera, typhoid, scarlet fever and syphilis, were brought under good control, or totally eradicated³. Sexually transmitted diseases (STDs) were so rare in China in the early 1980s that this topic was dropped from the medical school curriculum. The average life expectancy increased from 35 years in 1949 to 70 years in the early 1980s. Consequently, the Chinese experiment became the envy of the world, as no other large population had been so effectively, and so quickly, brought under the protective umbrella of a health-care system for all.

The present: tragic price

The transformation of China's economy from a 'social system' to a 'market-oriented system' began in 1978. Since then, China



has achieved remarkable economic development. The total poverty-stricken population, as defined by the government, decreased dramatically from 250 million in 1978 to 42 million in 1998. Yet the disparity between the rich and the poor has widened and despite ongoing economic growth, total government expenditure on public health has remained at the very low level of 5% of GDP (see graph, overleaf)^{4,5}.

Worse still, although spending on health, (as a percentage of total government expenditure) decreased from 28% to 14% between 1981 and 1993, allocation to the rural 'cooperative medical-insurance system' decreased from 20% to 2% (ref. 6). The health system has essentially collapsed in rural areas, and lack of financial support has forced many doctors out of health care altogether or into private practice, working on a fee-for-service basis. Low-income families in urban areas are now not covered by any insurance. The focus of 'prevention first' campaigns has been diverted. Infectious diseases that had been brought under control, such as tuberculosis and STDs, have re-emerged. Furthermore,

HIV/AIDS seems to have reached epidemic status in parts of China, although good epidemiological data are lacking.

In some rural areas HIV/AIDS is common because poverty forced many villagers to sell their blood to make a living, and they were infected from the replacement of mixed blood products and contaminated needles used during collection. About 800,000–1,000,000 HIV-positive people are now believed to live in the most poverty stricken rural areas of China. In addition to HIV/AIDS, hepatitis and liver cancer are also prevalent, and the scant attention paid to them increases their burden and toll. Because many low-income families have no health insurance or resources to purchase expensive medicines, such diseases are a premature death sentence.

Conflict of interest among doctors, hospitals and patients makes the situation even worse. Clinics and hospitals often prescribe excessive amounts of drugs, which are sold at high prices — a practice that has increased healthcare costs tremendously in the past decade. The corrupt practice of collecting a personal bribe or 'little red bag' has become routine among many doctors, especially in larger hospitals. Therefore Chinese citizens in most rural areas and low-income families in urban areas have difficulty gaining access to a doctor, are 'afraid' to get ill, or just do not seek treatment, even if they have a serious disease.

Contributing to this worsening picture are unhealthy lifestyles in China, exemplified by heavy smoking and poor diets. Such habits, together with bad industrial planning, such as siting residential housing next to substantial environmental pollution, have contributed to a dramatic rise in chronic diseases, including cancer, cardiovascular disease, hypertension and diabetes⁷.

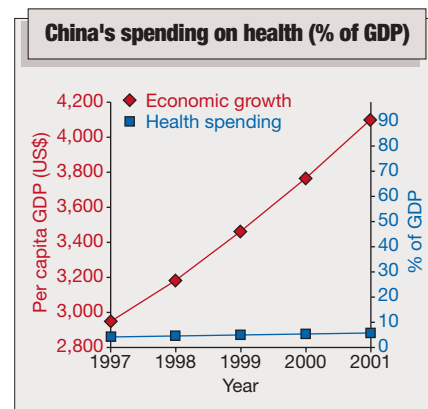
Finally, most high-quality hospitals, health care facilities and staff are located in urban areas and focus on disease treatment rather than prevention. According to the *World Health Report 2000* (ref. 8) the rural poor usually cannot receive quality health care, and in 'fairness of financial contribution' in terms of health spending relative to income, China was ranked 188 out of 191 countries. Corruption plays a big part in creating these inequities. In 2003, the Corruption Perception Index published by Transparency International placed China 66 out of 133 countries⁹. Today, corruption is found not only in health systems, but has become a major 'social disease' in China.

The future: rebuilding

There are not many reasons to admire Mao's economic policy, but his slogan "Give priority to prevention and provision of medical and health care in rural areas" was correct then and remains useful today. Rebuilding a primary health system for the 800 million people who live in rural areas

should be an immediate priority for the Chinese government, and there are signs that top government leaders are finally recognizing this. For example, in 2004 the People's Congress made 'three rural matters' (rural areas, rural agriculture, and rural populations) the country's highest priority, with health care being an important component in each. And in December 2004, the Ministry of Health announced a scheme in which, during the next three years, 10,000 urban doctors will be sent to rural areas to help with training and medical services⁵.

These plans will build on new strategies, outlined by the ministry in 2004, which include developing a response system for



emerging public health events and emphasizing the control of major infectious diseases, as well as establishing a "new type of cooperative medical care system"⁵. In the past year, around 333 counties in 30 provinces, with rural populations of about 107 million, joined the new insurance system⁵. In this, the central government contributes 10 Chinese Yuan (US\$1.25) for each person a year, with another 10 yuan requested from the local government and from individual farmers. The government hopes this system will go countrywide by 2010. Unfortunately, few local governments in the poorest areas, or rural populations themselves, can afford their share of the costs. According to a 2004 Ministry of Health report⁵, some local governments have simply increased general local 'taxes', or made fraudulent claims and falsified the number of participants. Such practices make the future of the project uncertain.

Since 2004, hepatitis-B vaccines became available to newborn children at no cost through a national vaccine programme. Considering that China has the largest hepatitis-B-infected population in the world (approximately 100 million people), such a prevention programme should soon produce results. But if the Chinese government is serious about addressing major infectious diseases — such as HIV/AIDS, tuberculosis and schistosomiasis, it needs to introduce more 'prevention first' policies.

As much as we welcome these changes, more is needed. Although the government is now increasing its spending on public health, the amount is still too small to meet the needs of 1.3 billion people. According to the World Health Organization, the per capita government expenditure on health in China was only \$18 in 2001 (ref. 4). This number is significantly lower than those of other developing countries with comparable incomes during the same period of time, such as Argentina at \$679, Chile at \$303, Brazil at \$222, Cuba at \$185 and Malaysia at \$77. Continuing to dramatically increase investment in public health is critical for China's continued economic growth and social stability.

Reform from within

One cannot expect China's public health system to thrive without simultaneously conducting an anticorruption campaign that prevents the over-pricing of medical care and pharmaceuticals, while guarding against preferential treatment for those with better finances and a 'little red bag'. Medical ethics in China is in its infancy, but training medical students, clinicians, public-health officers and government officials in the fundamentals of medical ethics is essential for successful reconstruction of the system. The Chinese government is finally starting to pay attention to this problem. In 2004 they reported that 229 doctors and 1,587 people were 'punished' for collecting 'red bags' and overpricing medical care, respectively⁵. Sadly, corruption exists in almost every professional activity in China, as well as most levels of government. An effective campaign to end these practices must extend to fields beyond health care.

None of these goals is easily accomplished. But the key to solving the immediate problems of China's public health is now in the hands of the Chinese government. Ignoring the public-health crisis will eventually lead to other socio-political problems and will ultimately limit economic growth. Initiating a comprehensive, prevention-focused programme today will mean a healthy and prosperous China tomorrow. ■

Zigang Dong is at the Hormel Institute, University of Minnesota, Austin, Minnesota 55912, USA.

Christina W. Hoven and Allan Rosenfield are at the Mailman School of Public Health, Columbia University, New York, New York 10032, USA.

1. Sidel, V. & Sidel R. *Soc. Sci. Med.* **11**, 415–419 (1977).
2. Sidel, R., Sidel, V. & Sidel, M. *Health of China* (Beacon Press, Boston, 1982).
3. Rosenfield, A. G. *Man Med.* **2**, 279–296 (1977).
4. World Health Organization *The World Health Report 2004 — Changing History*; available at www.who.int/whr/2004/en
5. Ministry of Health of the People's Republic of China; available at www.moh.gov.cn (in Chinese)
6. Bloom, G. & Gu, X. *Soc. Sci. Med.* **45**, 351–360 (1997).
7. Lee, L. *Annu. Rev. Pub. Health* **25**, 327–339 (2004).
8. World Health Organization *The World Health Report 2000 — Health Systems, Improving Performance*; available at www.who.int/whr/2000/en
9. *Global Corruption Report 2003*; available at www.globalcorruptionreport.org

Feet of clay

There's more to science than doing the research.

Leaps in the Dark: The Making of Scientific Reputations

by John Waller

Oxford University Press: 2004. 304 pp.

£18.99, \$24.95

Walter Gratzer

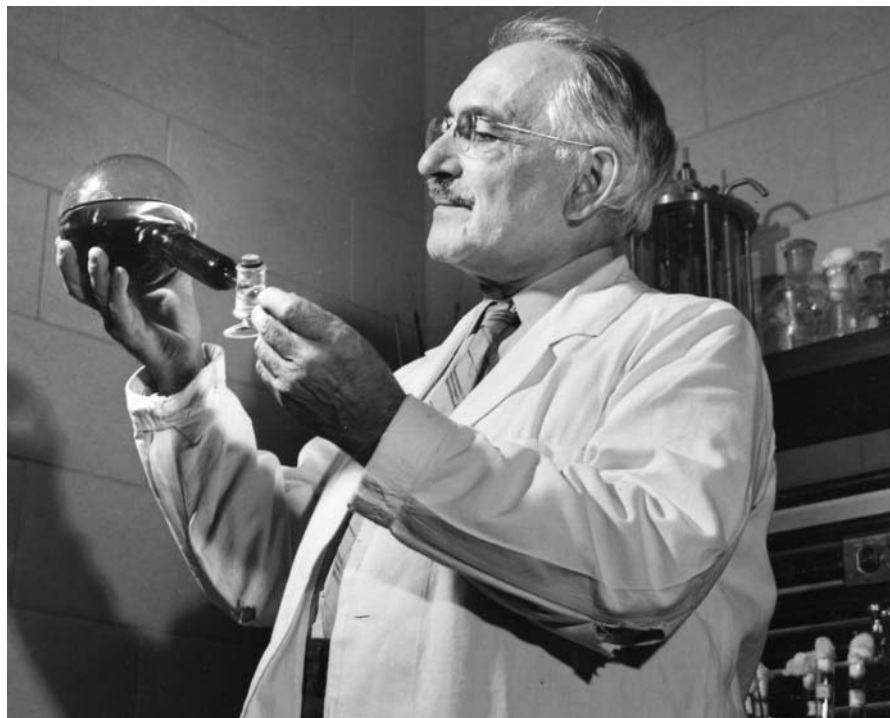
"The next day Koch was bending nearsightedly over the body of this little creature on his dissecting board; giddy with hope, he was carefully flaming his knives... Not three minutes later Koch is seated before his microscope, a bit of the dead creature's spleen between two thin bits of glass. 'I've proved it,' he whispers, 'here are the threads, the rods — those little bacilli from my hanging drop were just as murderous as the ones right out of the spleen of a dead sheep.'" Such was the style of popular science writing in the dawn of the genre some eighty years ago. Paul de Kruif's reams of incandescent prose, hard to stomach today, inspired two generations of schoolchildren to follow in the footsteps of Robert Koch and other heroes.

John Waller finds such romantic images of the scientific process pernicious, and he is a man with a mission: he wants to rip the laurels from the brows of the scientific conquerors, and to place their achievements in a more realistic, even at times sordid, context. *Leaps in the Dark* is his second foray into this territory, and follows on from *Fabulous Science*.

We know of course that no important discovery ever emerges from an intellectual vacuum, for it is a truism that every advance rests on what has gone before. Even Newton, Waller reminds us, admitted to having seen further than his peers because he had "stood on the shoulders of giants" — though the remark is generally interpreted as a dig at his detested rival, the diminutive Robert Hooke, rather than a mark of modesty, an attribute wholly alien to Newton's temperament.

Waller excoriates the sin of 'presentism' — the interpretation of historical events in terms of the social, moral and intellectual norms of our own time. So, he avers, it was the intellectual constraints of the period that drove the great biologist and polymath Lazzaro Spallanzani to reject the (to us self-evident) implication of his own remarkable experiments: that an animal's genetic composition does not derive solely from its mother.

The attempt to supplant an existing principle that has until then adequately accommodated the known facts inevitably meets resistance. But this reaction, as Waller stresses, must not be taken to imply that the



Unprincipled investigator? Selman Waksman took all the credit for the discovery of streptomycin.

opposition is foolish or bigoted. An example is the power that the germ theory of disease exercised in the late nineteenth century, which made it hard for scientists and doctors to get their minds around the idea that disease could equally be caused by the absence of something (such as trace nutrients).

But with his relentless emphasis on the influence of social and political factors in determining whether a new theory prevailed or sank, Waller erects some speculative structures of his own. He holds, for instance, that in the prolonged debate over the cause of cholera, Koch vanquished Max von Pettenkofer not because he was right and his adversary was wrong, but because the Kaiser and the Chancellor of a newly united Germany were eager to present the fiercely patriotic and gallophobic Koch as their country's rival to Louis Pasteur and the French school.

And how did Ignaz Semmelweis's insight that childbed fever was disseminated by a contagion on the hands of doctors and students coming from the autopsy room to the labour wards fail so signally to find favour with his superior, Johannes Klein? Why, because of Klein's distaste for Semmelweis's republican politics, of course, rather than for his theory. Well, maybe.

In Waller's summing-up we find the following passage: "By the end of the 1800s scientific advance remained closely linked to political and economic developments. As

the efficient harnessing of power became of vital economic importance with the onset of industrialization, physicists turned from the study of matter to that of energy." Leaving aside the implicit conflation of science with technology, is there evidence that the likes of Hermann von Helmholtz, Julius Mayer, Rudolf Clausius and Ludwig Boltzmann concerned themselves greatly with economic imperatives? And did Darwin really draw his inspiration from the struggle for existence in the mill towns of the Industrial Revolution? There must at least be scope for doubt. Waller's conjectures are certainly not without interest, and he seldom fails to make a lucid case. But he threatens, it seems to me, to replace one form of historical distortion (Thomas Carlyle's 'great man' view of human progress) with another.

Waller concludes with two cases of misrepresentation and injustice that are still within living memory. He does a magisterial demolition job on the egregious scientific mandarin Robert Watson-Watt, whose outrageous claims to have invented radar and directed its operation were scarcely questioned by the politicians. R. V. Jones, in his wartime memoirs, recalls his efforts to escape from Watson-Watt's capacious orbit, having found him to be devious, manipulative and, worse, a third-rate physicist. Watson-Watt's boundless self-aggrandisement bore fruit, and his is still the name most often

associated with the invention that, perhaps more than any other, secured victory for the Allies in the Second World War.

Selman Waksman, by contrast, was a well respected researcher who forsook his principles for the irresistible prospect of lasting fame. The antibiotic streptomycin eventually saved the lives of innumerable consumptives, and earned Waksman a Nobel prize in 1952. His claim to have discovered the antibiotic essentially by himself was challenged by his student Albert Schatz, who went so far as to take his supervisor to court, an act that incurred the opprobrium of the scientific establishment for bringing their calling into disrepute. The affair put paid to Schatz's career, even though it was he who had on his own initiative isolated streptomycin and tested it against tubercle bacilli. As so often in such cases, the student was convinced that the credit was entirely his, while the professor believed that he had created the intellectual milieu which opened the way to the discovery. This does not exculpate Waksman, who had, as Waller shows, fallen into the grip of self-delusion. Friedrich Nietzsche explained this psychology most concisely: "I have done this," says my memory. "I could not have done this," says my pride, and remains adamant. At last memory yields."

Let me affirm finally that *Leaps in the Dark* is a good read, and ought to generate much healthy debate. Waller has opened what seems to be an inexhaustible vein — one that will no doubt yield more high-quality ore. ■

Walter Gratzer is at the Randall Division of Cell and Molecular Physics, King's College, New Hunt's House, Guy's Campus, St Thomas' Street, London SE1 1UL, UK.

Poison in the well

Venomous Earth: The World's Worst Mass Poisoning

by Andrew A. Meharg

Macmillan Science: 2005. 256 pp. £16.99, \$29.95

Roger P. Smith

The dust-jacket shows cupped palms in a pose eerily reminiscent of that used by the US insurance company that bills itself: "The Good Hands People". The skin, however, is pigmented by the black raindrop pattern of melanosis characteristic of chronic arsenic poisoning. The victim is one of a predicted 350,000 people in Bangladesh and West Bengal, India, who will develop fatal cancers over the next ten years.

As one of the poorest nations in the world, Bangladesh has long lacked even necessities such as safe drinking water for its burgeoning population. Before the early



Arsenic in the drinking water leaves its mark on the hands.

1970s, a quarter of a million children died each year from water-borne communicable diseases such as cholera and dysentery. In what better hands to place this catastrophe than those of UNICEF, the United Nations children's fund?

Acting under the purest of humanitarian motives, UNICEF proposed a solution that seemed a providential blend of economy and speed. Inexpensive iron tubes were pushed through the marshy ground to tap water stores just beneath the surface. When fitted with hand pumps, an inexhaustible supply of clean water seemed assured. By 1972, nearly a million such devices had been installed, and the health benefits were already evident. The mortality rate for children under five began to drop and was halved by 1996. Efforts to provide the entire area with safe water were redoubled, and now an estimated 95% of rural inhabitants have access to at least one such well.

But in the background an unimagined horror was slowly emerging. The first confirmed report of arsenic contamination of tubewell water came as early as 1983, but it was regarded as an isolated incident. A dermatologist at the Calcutta School of Tropical Medicine recognized the skin lesions, and had some West Bengali wells investigated. They tested positive for arsenic. By 1987, more than 1,000 cases had been identified. But papers published in local and international journals were largely ignored. Organized conferences and pleas for attention fell on the deaf ears of aid officials and government workers, who pointed to the obvious

health benefits. The installation of the tubewells continued to accelerate. Cases of arsenic poisoning began to appear in Bangladesh, where the problem was found to be even more severe than in West Bengal.

How the water came to contain such high levels of arsenic is still a matter for conjecture. Why the distribution of arsenic contamination is so maddeningly random is still not known. The only timely recourse seemed to lie in testing each well and painting the pumps red if they were contaminated and green if they were safe. The water's arsenic content tends to change with the time and season, however. Much deeper tubewells were more likely to be safe but were more expensive, and replacing the hundreds of thousands of contaminated shallow tubewells could take years.

This tragic but important story is told in the first two chapters of this slim volume, with some solutions proposed in the final chapter. In between the book deals with the history of arsenic. There is a unique section on the commercial uses of arsenic pigments, such as Scheele's green and emerald green, which satisfied a void in the colour spectrum for consumer products such as wallpaper. The use of arsenical dyes in the nineteenth century spread to clothing, paper, cardboard, soap, toys, paints, artificial and dried flowers, stuffed animals, Venetian blinds, curtains and ballgowns. They were even used as food colourings. The morbidity and mortality that resulted from this collective insanity will never be fully tabulated.

There is a chapter on the largely ill-advised therapeutic uses of arsenic, and another on the homicidal application of arsenic, including some of the more famous poisoners and their victims. Also discussed is the ugly environmental impact of arsenic production or its release from the commercial smelting of ores for other materials.

There is certainly much here to fascinate a general audience but I found the book to be curiously disjointed. There is little mention of the biological mechanisms by which arsenic produces its effects, for example. Earth scientists may find enough to identify with, but perhaps the author could have been persuaded by his editor to provide a little more for life scientists. ■

Roger P. Smith is emeritus professor of pharmacology and toxicology, Dartmouth Medical School, Hanover, New Hampshire 03755, USA.

The force behind the prize

The Quantum Quark

by Andrew Watson
Cambridge University Press: 2004. 476 pp.
\$30, £19.99

John Ellis

The Nobel Prize for Physics in 2004 was awarded to David Gross, David Politzer and Frank Wilczek for their discovery in 1973 of 'asymptotic freedom', the key ingredient in the underlying theory of the strong nuclear force. Journalists have had a hard time trying to explain this abstruse breakthrough to the public. Andrew Watson's book *The Quantum Quark* is probably not cut out for coffee-table stardom, but it is a very complete account of the ideas and experiments that led to the development of this theory and its verification. As such, it may appeal to anybody willing to take the time and make the effort to understand the basis for last year's prize.

Many of the 'usual suspects' of twentieth-century physics parade through the pages of this book: quantum mechanics, relativity, quantum field theory, and so on. But they are marching towards a specific goal, namely their roles in quantum chromodynamics (QCD), the underlying theory of the strong interactions. It is not every day, nor even every century, that a new layer of the 'cosmic onion' is revealed, so the discovery of the underlying quark structure of nuclear particles and the quantitative understanding of the strong interactions must rank among the highest intellectual achievements of twentieth-century physics.

However, QCD is not a glamorous theory and has not made as big a splash as the unified theory of the weak interactions, with its flashy discoveries of the W and Z particles and the long anticipation of the enigmatic Higgs boson. Nevertheless, QCD has great inner beauty, is just as much a part of the standard model of elementary particles, and was from some points of view a tougher theoretical nut to crack.

Watson certainly gives us plenty of theoretical red meat to sink our teeth into, perhaps more than some delicate stomachs can accommodate. As well as being very complete, this book is also reasonably accurate. However, I did notice one scientific howler: electromagnetism is not described simply by the U(1) part of the electroweak theory, but is a more complicated mixture of its different components. Also, I am concerned that the lay reader might find its early sections challenging: the explanations are reasoned and self-contained, but veer towards the succinct.

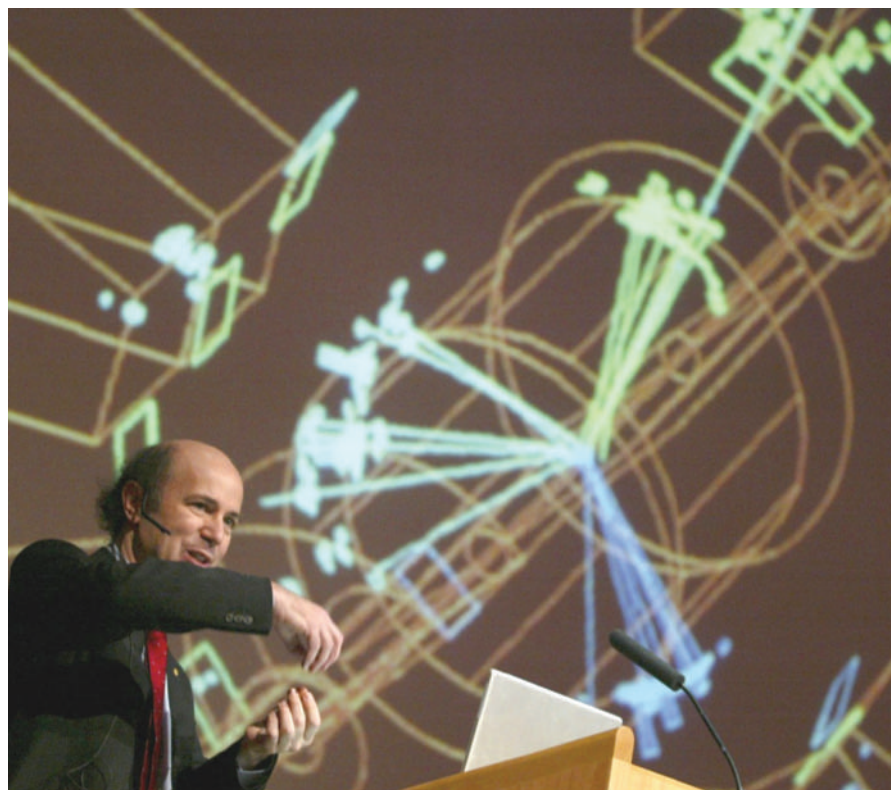
Personally, I would also have preferred more experimental sauce in this early part.

After all, experiment is the final arbiter of whether nature has the good taste to follow our theoretical ideas. It might have been good, for example, to explain what it was about the photoelectric effect that led Einstein to postulate the existence of the photon, the prototype force-carrying boson that was joined later by the gluons that bind the quarks, and by the W and Z particles responsible for the weak interactions. Nevertheless, this introduction and review of basic ideas in field theory and particle physics makes a good appetizer.

Unfortunately, I found the main course unsatisfactory in several respects. First, the scientific events are often not presented in

These topics are all discussed, but not with as much attention and emphasis as I should have liked. Conversely, some more controversial aspects of QCD, such as the everlasting hunts for glueballs and exotic bound states, receive more attention than they may yet deserve. Another gripe is that the book is probably too compendious for many readers, who may get lost in a forest of miscellaneous tree diagrams. I think it would have benefited from more active editing.

So do I recommend this book? The author and his publishers are certainly to be congratulated on their uncanny timing. It is a real coup to have released this book just at the moment when physics students, scientists in



Attractive idea: Frank Wilczek won a Nobel prize for work on the theory of the strong interaction.

chronological order. This might sometimes be a useful literary device, but why discuss deep-inelastic scattering experiments before the earlier pioneering ones that revealed quark partons, or the even earlier ones demonstrating that the proton was not a simple point-like object? There are also some irritating factual errors in this more experimental part. For example, the lower limit on the lifetime of a proton is 10^{29} years rather than 10^{25} years, the lower limit on the mass of the Higgs boson is 114 GeV rather than 65 GeV, and the maximum energy achieved by the Large Electron-Positron accelerator at CERN was 208 GeV rather than 192 GeV.

A more basic shortcoming of the book, in my opinion, is its failure to bring out some of the beauties of QCD, such as the renormalization group and asymptotic freedom, or confinement and the nature of the vacuum.

other fields, and general readers around the planet may be keen to unravel for themselves the context for this year's physics Nobel. Given this motivation, they may find this book an informative and fascinating read.

Moreover, although it has some flaws, this book does not have much competition. Andrew Pickering's *Constructing Quarks* (Edinburgh University Press, 1984) is a real page-turner and is scientifically accurate, but is marred by its aberrant initial and final chapters full of sociological mumbo-jumbo, and is now rather dated. There is still scope for someone to write the definitive book on quarks and QCD, but in the meantime, I suggest that you persuade your library to buy Watson's book — although you might want to think twice before buying it yourself. ■

John Ellis is in the Theoretical Physics Division, CERN, Geneva 23 CH-1211, Switzerland.

Schrödinger's mousetrap

Part 4: A very public humiliation.

Laura Garwin

Nigel Lorimer was worried. He'd seen enough television detective shows to suspect that the dishevelled man opposite him, who was currently fighting an unequal battle with a packet of nicotine gum, probably had a mind like a steel trap. "Come on, Nige, relax," he said to himself. "If you can hold your own in a dispute with an aggrieved member of the National Academy of Sciences, you can certainly handle this flatfoot."

With the gum finally in his mouth and the drug coursing reassuringly through his veins, Lister returned to the question at hand: "And where were you between 10:30 and 11:00?"

Lorimer thought for a moment. "Let's see, I ducked out of the session on colossal magneto-resistance at about 10:30, made a quick phone call back to the office, and then spent the rest of the time before the lecture in the hallway by the coffee and biscuits."

"Did you speak to anyone in the coffee break?"

"Did I speak to anyone?! What do you think *Nature* pays me for? Plus there are always people wanting to speak to me."

Lister asked for details, and Lorimer duly supplied the names of several physicists who could vouch for his presence. "I don't suppose you spoke to Professor Jaeger?" Lister asked.

Lorimer flushed at the sound of the dead man's name. "No," he said, after a short pause. "Rufus was busy having some kind of dust-up with that sidekick of his, Wilfred de Bruijn."

Now it was Lister's turn to pause. "Are you sure that's why you didn't speak to him?" he said, locking his eyes on Lorimer's. "From what I hear, the professor wasn't exactly in your good books."

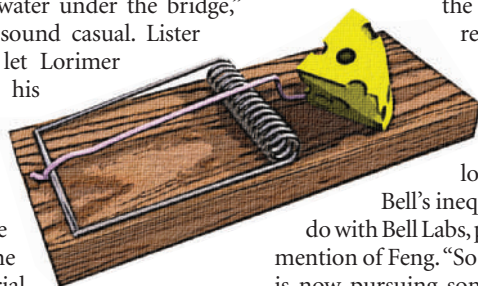
Despite himself, Lorimer began to sweat. "If you're referring to the infamous quantum-relay paper, that's water under the bridge," he said, trying to sound casual. Lister said nothing, but let Lorimer tell the story at his own pace, which accelerated as his thoughts returned to that embarrassing time. For, despite his protestations, the wound to his editorial pride was still raw.



The paper had looked so good at the start. Even now, Lorimer could remember the little shiver of excitement he'd felt when he'd read it for the first time — and how grateful he'd been to Jaeger for choosing to send the paper to *Nature*. The referees were equally enthusiastic — one had called the work a landmark in the field of quantum information — and the paper appeared to great acclaim in the first issue of 1999. But then it all went horribly wrong. First there were whispers in the corridors at conferences, then letters to the editor pointing out inconsistencies in the results, and finally the retraction, published towards the end of the same year.

In public, as would be expected, all of the authors shared the blame for the miscalibrated detector that led to the rogue result. But everyone assumed that the real culprit was the paper's first author, Jaeger's brilliant but erratic graduate student Jirong Feng. After all, Feng had been dismissed from Jaeger's group shortly before the retraction was published. "I've always wondered why our referees — one of them, in particular — didn't spot the mistake during the review process," Lorimer concluded, with a sigh.

Lister, who had been wondering what an entangled photon looked like, and whether Bell's inequalities had anything to do with Bell Labs, pricked up his ears at the mention of Feng. "So I suppose this guy Feng is now pursuing some less desirable career track — investment banking, perhaps?"



"You would think so, wouldn't you?" Lorimer replied. "But that's the odd thing. A year or so after Jirong left Rufus's group, Petra Pruszczyński took him on at Gdansk. I think he's nearly finished his PhD now, and he seems to be doing some great work."

"Really? What's he working on?" asked Lister, writing something down in his notebook.

"You can ask him yourself," Lorimer replied. "He's here at the conference."

"And this Professor Prusz... Pruszyk... you know, the woman in Gdansk — have you had any dealings with her lately?"

After a moment's hesitation, Lorimer replied, looking pleased with himself. "Well, just between you and me, I'm handling a paper from her at the moment — one that could be quite important, if it

stands up."

"And who are the referees of that paper?" Lister asked.

Lorimer bristled. "Now you're going too far. That's deeply confidential information. And besides, what possible reason could you have for wanting to know?"

Lister responded gravely, once again holding Lorimer's gaze with his own. "Dr Lorimer, suffice it to say that I have reason to believe there is a connection between Professor P's paper and Professor Jaeger's murder."

Lorimer caught his breath. "All right," he said slowly, "I'll give you the names of the referees. But you must keep them to yourself." Lister nodded, holding his pen at the ready. "One of them is Fenton Baumgarden, who's here at the conference. The other was Jonas Prirsali from Innsbruck."

"Was?" Lister asked sharply, looking up from his notebook.

"Yes. Sadly Jonas died two weeks ago from a heart attack. He must have been in his seventies, and he hadn't been well for quite some time. But at least he finished his referee's report before he died."

"You sound like a very dedicated editor," Lister said drily, as Lorimer's ears turned pink with embarrassment.

To be continued...

Laura Garwin is executive director of the Bauer Center for Genomics Research, Harvard University, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA.

Missed an episode? Catch the story so far at www.nature.com/news/mousetrap

CHRISTIAN DARKIN

CHRISTIAN DARKIN

The onset of selection

Natural selection started to drive evolution as soon as molecular replication became possible.

Christian de Duve

Divined by the genius of Alfred Russell Wallace and Charles Darwin, the basic principles of evolution by natural selection are well known. First, there is genetic continuity, based on replication. Then, inevitably, comes variation. Finally, there is competition, leading to selection of the variants most apt to survive and proliferate under the prevailing conditions. The findings of modern biology have fully validated those principles, adding the fundamental fact that the causes of variation are strictly accidental and unintentional.

The key notion in this theory is replication. The rest follows obligatorily. Thus, in the origin of life, darwinian evolution must have started as soon as the first replicable molecules appeared. Here, I intend to draw attention to certain implications of this fact that are sometimes ignored or underestimated in discussions of the origin of life. I shall assume, in agreement with most workers in the field, that the first replicable molecules consisted of RNA.

Direct versus indirect selection

The simplest manifestation of natural selection is the direct, molecular form — the object of many studies since it was first made to occur in the test tube by Sol Spiegelman in the 1960s. If RNA molecules are allowed to replicate *in vitro*, selection automatically screens out those mutant molecules that best combine stability and replicability — the molecular equivalent of darwinian survival and proliferation — under the adopted conditions.

By necessity, this kind of selection must have started with replication. In fact, the first product of molecular selection may well have been RNA itself. The mechanism whereby this substance arose is still unknown, but cannot possibly, unless guided by some prescient agency, have produced only authentic RNA molecules with the bases A, U, G and C as sole constituents. It is much more likely that such molecules were accompanied by other analogous assemblages and that they were selected out of this mixture and amplified, thanks to their ability to induce, by base pairing, the formation of complementary molecules that could in turn act similarly to reproduce the first ones.

Once initiated, such a process would have evolved naturally toward the production of what Manfred Eigen has called the Ur-Gen, the ancestor of all RNA molecules. This product would have arisen by molecular selection to form a 'quasi-species', consisting of a 'master sequence', optimized with respect to the prevailing environment, and of an ever-changing cohort of mutants arising through replication errors and other accidents.

The RNA molecules that initiated protein synthesis probably belonged to this early

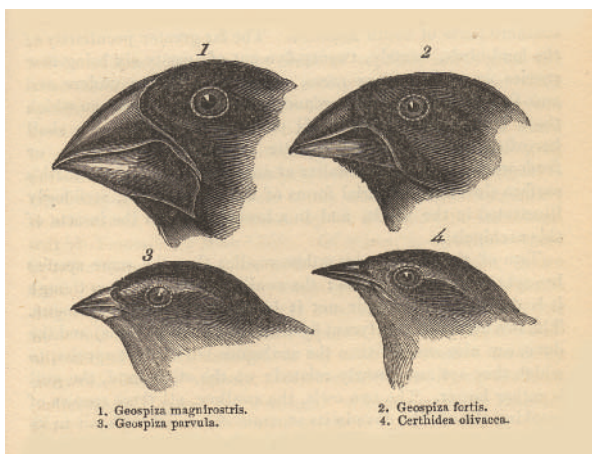
been reached where direct selection ceased to be the sole operating process and a new, indirect form of selection was initiated because of the growing complexity of the system.

In this form of selection — which dominates darwinian evolution — genes are selected not because of what they are, but because of what they or their products do, which in the beginning must have been mainly to catalyse a chemical reaction useful to the gene's replication. Barring rare exceptions (such as self-replication), the criterion of usefulness requires the reproduction of the gene to be linked to that of an entity that derives an advantage from the new reaction allowed by the gene. This condition almost mandates the existence of primitive cells, or 'protocells', able to grow, multiply and compete with other protocells for available resources.

It follows that cellularization must have occurred very early in the development of life, probably no later than the inception of protein synthesis. Most of the catalytic RNAs (ribozymes) involved in the so-called 'RNA world' and all the first protein enzymes must have been 'invented' by protocells capable of participating in darwinian competition and of deriving a selective advantage from the catalyst. An important implication of this conclusion is that the early protocells must have been sufficiently individualized to be able to engage in competition and benefit from selection. This point is relevant to the proposal, advocated by Carl Woese, W. Ford Doolittle and others, that the first cells, up to the last universal common ancestor (LUCA) of all known living beings, formed a collective of ill-defined entities that freely exchanged and shared genes by horizontal transfer. This phenomenon may, indeed, have been more important in the early days of the development of life than it is today — which is not inconsiderable — but it cannot have been so extensive and frequent as to blur the distinctions between individual lineages, suppress competition and impede selection.

Mutual selection

Selection is usually visualized as a one-way process, in which a shifting collection of evolving entities is subject to screening by the environment. But the process is often mutual, with the environment being itself screened by the evolving entities. This reciprocity is asymmetric, involving an active and a passive



The Galapagos finches inspired Darwin's theory of natural selection.

crop. It is widely believed that protein synthesis was launched by interactions between RNA and amino acid molecules, prefiguring the present role of transfer RNAs as carriers of the amino acids that are incorporated into proteins. Whereas the amino acids were present beforehand, the RNA partners of the primeval associations most likely arose as mutants of the Ur-Gen and were subsequently selected. This could have happened by a molecular mechanism. An RNA molecule bearing an amino acid could have adopted a more stable conformation. More importantly, for a long-lasting effect, it could have interacted more efficiently with the replication catalyst, thus furthering its own replication.

Other RNA molecules presumably also participated in the development of protein synthesis, for example by favouring the proper alignment of RNAs bearing amino acids, or by catalysing peptide bond formation, functions fulfilled today by messenger RNAs and ribosomal RNAs, respectively. Selection of such RNAs by a molecular mechanism cannot be ruled out, but is not readily visualized. In any case, a stage in the development of protein synthesis must have

partner. The active partner is a replicable entity subject to genetic variation and natural selection, whereas the passive partner is no more than a component of an existing pool, from which it is simply singled out.

Illustrating such an occurrence is the mechanism previously outlined, whereby certain RNA molecules (the active partner) may have been selected by virtue of their ability to combine with amino acids (the passive partner). If, as seems likely, there was any chemical specificity in the interactions involved, the selected RNA molecules must have acted reciprocally to recruit out of the prebiotic pool the first amino acid molecules that came to be used for protein synthesis — thereby launching the selection of the amino acids, including their chiral forms, that universally serve for the construction of proteins.

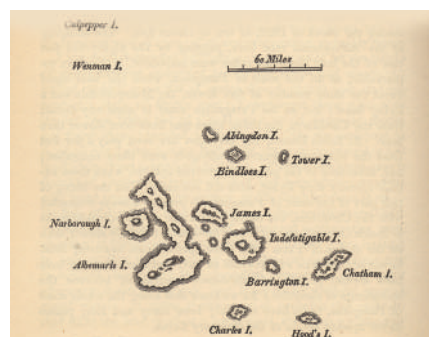
The development of metabolism offers a second possible instance of mutual selection. Here, the active partners were either RNA ribozymes or RNA-encoded protein enzymes. The passive partners belonged to the pre-existing chemistry (protometabolism), which must have supplied the catalysts with one or more substrates to act on and with one or more outlets for the product(s) of the catalysed reactions — otherwise, the catalysts could not have been selected. Without substrate or outlet, a catalyst cannot be useful; it may even be harmful. This relationship, which is rarely appreciated, implies that some of the key features of present-day metabolism must have been prefigured in protometabolism; the two were congruent. A corollary is that the catalysts acted reciprocally to select out of protometabolism reactions (or, at least reactants) that came to be taken up into metabolism. In other words, certain components of early metabolism were selected from what was perforce a 'dirty' protometabolism by the catalysts they themselves served to select.

Optimizing selection

A common opinion is that the course of evolution, being dependent on chance occurrences, is utterly contingent, unpredictable and non-reproducible. This is not necessarily so; chance does not exclude inevitability. It all depends on the number of opportunities provided for an event to happen, relative to the event's probability. Even a seven-digit lottery number has a 99.9% likelihood of coming out if 69 million draws are held. Lotteries don't operate that way, of course, but the evolutionary lottery often does, thanks to an extensive, if not exhaustive, exploration of the possibilities open to a system.

An easily quantifiable case is represented by the sequence spaces of nucleic acids and proteins, that is, the number of distinct sequences of a given length that are possible. A simple calculation shows that, for molecules

of the size found in living organisms today, the sequence spaces are unimaginably immense, and the region occupied within the spaces by biomolecules is vanishingly small. In the eyes of the small but vocal minority of scientists who defend 'intelligent design', life could not have reached this position by chance alone; it must have been 'guided'. Another less controversial view is that the sequence spaces contain enough life-compatible 'islands' for one such island to be reached with a fair degree of probability by a random walk, but by a pathway that



The Galapagos Islands, as drawn in 1835.

is utterly contingent, unpredictable and non-reproducible, in agreement with the prevailing doctrine.

Such views rest on false premises. Life did not start with the kind of large nucleic-acid and protein molecules it uses today. Molecules of that size (and even much smaller ones) would, as demonstrated by Eigen, inevitably have degenerated upon repeated replication, due to the cumulative errors caused by the necessarily low fidelity of the primitive RNA-replicating systems. For this reason, the first replicable RNA molecules and their protein translation products must have been very short, and their subsequent lengthening — probably by splicing of pre-existing RNA stretches — was imperatively dependent on the development of RNA-replicating systems of greater fidelity. Protein molecules thus most likely reached their present size by a stepwise process, 'bootstrapped' at each stage by some improvement in the fidelity of RNA replication, presumably made possible by the appearance of longer, more complex RNA or protein molecules to serve as catalysts.

An important implication of the necessarily stepwise course of RNA and protein lengthening is that exploration of the corresponding sequence spaces was itself stepwise. It may have started with a small enough number of possibilities (short enough stretches) to allow selection to produce the molecules best adapted to the prevailing conditions, by way of the protocells. The number of such molecules could, in turn, have been small enough to allow selective optimization at the next step, and so on. In other words, selection could, at each size level, have reduced the

number of molecules available for splicing down to a value compatible with exhaustive or near-exhaustive exploration of combinatorial possibilities, thus approaching optimization at each stage. It follows that nascent life could have reached the infinitesimally minute area it occupies in sequence spaces by a succession of selective bottlenecks, so that the pathway ended up being close to obligatory, at least in its main lines, under the prevailing environmental conditions.

Another instance of optimization may be represented by the genetic code. According to computer-simulation experiments by Stephen Freeland and colleagues, the code seems to be to be more efficient than average — by approximately six orders of magnitude — in minimizing the changes in hydrophobicity, and thus presumably, the harmful consequences resulting from point mutations leading to the replacement of one amino acid by another. If this quality of the genetic code should indeed be a product of natural selection, it would represent a particularly impressive case of optimizing selection, as it implies that primitive protocells had the opportunity to 'experiment' with a large number of distinct genetic codes, the selective criterion being the progeny's long-term ability to withstand the consequences of point mutations. The conditions that would have allowed this kind of experimentation raise challenging questions.

Even later evolution could have enjoyed less freedom than is often claimed, as indicated by phenomena such as drug resistance, mimicry and convergent evolution. It seems that, in many cases, evolving systems had the opportunity to put to the test a vast enough array of relevant mutations to evolve near-optimal, reproducible solutions to environmentally imposed survival problems.

In conclusion, darwinian selection started operating very early in the origin of life and probably played a major role in the shaping of the first living cells.

Christian de Duve is at the ICP, B-1200 Brussels, Belgium, and at the Rockefeller University, New York, New York 10021, USA.

FURTHER READING

- de Duve, C. *Life Evolving* (Oxford Univ. Press, 2002).
- de Duve, C. *Singularities* (Cambridge Univ. Press, 2005).
- Eigen, M. & Winkler-Oswatitsch, R. *Naturwissenschaften* **68**, 282–292 (1981).
- Eigen, M. & Schuster, P. *Naturwissenschaften* **64**, 541–565 (1977).
- Woese, C. R. *Proc. Natl Acad. Sci. USA* **95**, 6854–6859 (1998).
- Doolittle, W. *Curr. Opin. Struct. Biol.* **10**, 355–358 (2000).
- Freeland, S. J. *et al. Orig. Life Evol. Biosph.* **33**, 457–477 (2003).
- Conway Morris, S. *Life's Solution* (Cambridge Univ. Press, 2003).

Acknowledgements

I am greatly indebted to Gerald Joyce for valuable comments on an early draft of this essay. Neil Patterson has contributed insightful editorial suggestions.

Gene exchange by design

Stanton B. Gelvin

Gene transfer from bacteria to plants was thought to be limited to the bacterial genus *Agrobacterium*. But other bacterial groups also contain species capable of interkingdom genetic exchange.

All organisms must be able to transfer genetic information from parent to progeny — without this 'vertical' gene transmission, species would simply die out. But this is not the only mechanism for transmitting genes from one organism to another. 'Horizontal' gene transfer has long been recognized as a major factor in evolution, and there are numerous likely examples of past horizontal gene transfer events. In terms of ongoing processes, bacterial conjugation (gene transmission from one bacterium to another through direct physical contact) is the most commonly cited example. The horizontal transfer of genetic material between species from different phylogenetic kingdoms — from bacteria to plants, for instance — is also possible, but had previously been documented only for members of the bacterial genus *Agrobacterium*. Now, on page 629 of this issue, Broothaerts *et al.*¹ show that several bacterial species outside this genus are also capable of interkingdom 'sex'.

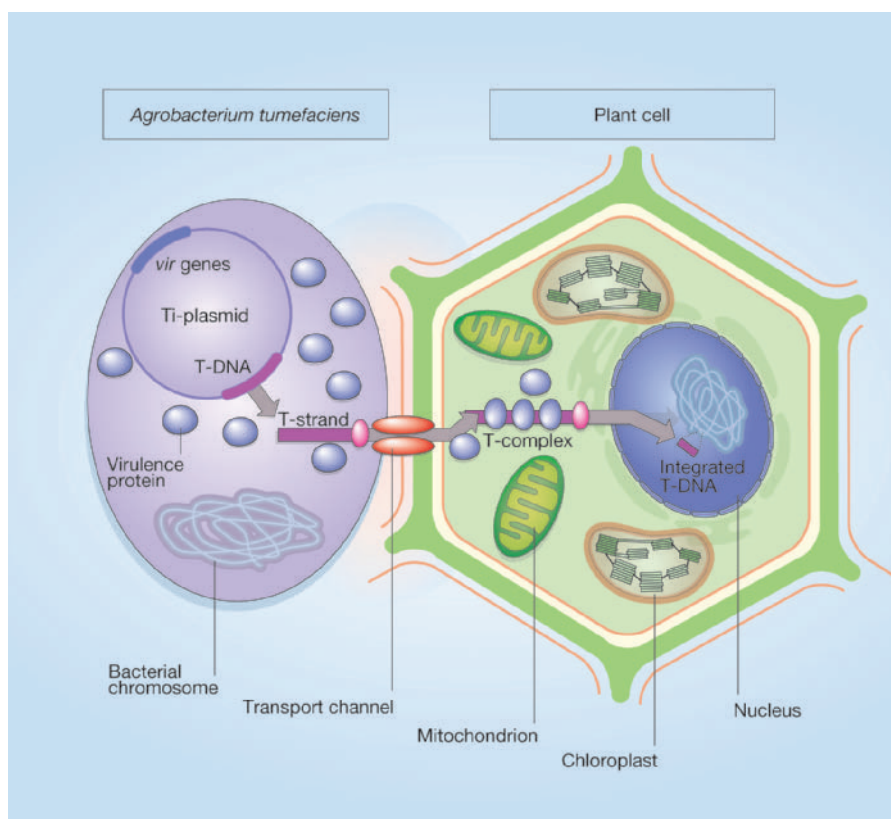
Agrobacterium tumefaciens is a pathogenic bacterium that causes crown gall disease (Fig. 1), a cancerous disorder that affects a range of plant species². Along with the related species *A. rhizogenes*, *A. vitis* and *A. rubi*, *A. tumefaciens* incites tumours by transferring a portion of a large extrachromosomal DNA element (the tumour-inducing, or Ti, plasmid) to its plant hosts. Once within the plant cell, this transferred DNA (the T-DNA) targets the nucleus, where it eventually integrates into the host genome. The expression of T-DNA-encoded oncogenes results in the plant overproducing or becoming overly sensitive to hormones, leading to loss of cell-growth control and to tumour formation³.

This ability of *Agrobacterium* species to transfer DNA to plants has been exploited by scientists for genetic-engineering purposes; oncogenes can be deleted from the bacterial T-DNA and substituted with any gene of interest (Fig. 2). Once transferred, these 'transgenes' can endow plants with new characteristics, including herbicide or pathogen resistance, altered growth or nutritional qualities, or the ability to produce drugs or edible vaccines — antigens that can direct immunity against pathogens in humans or animals that ingest the plants. Under laboratory conditions, *Agrobacterium* can also genetically transform fungal and animal cells^{4–6}.



Figure 1 Crown gall tumour on an elm tree. Most crown gall tumours form at or below the surface of the soil, but aerial tumours can also occur, as here. (Courtesy of Jer-Ming Hu, National Taiwan University, Taipei.)

Figure 2 How *Agrobacterium* genetically transforms plants. *Agrobacterium* contains a tumour-inducing (Ti) plasmid, which includes virulence (*vir*) genes and a transferred-DNA (T-DNA) region. Genes of interest can be inserted into the T-DNA. Wounded plant cells produce phenolic defence compounds, which can trigger the expression of the *Agrobacterium* *vir* genes. The encoded virulence (Vir) proteins process the T-DNA region from the Ti-plasmid, producing a 'T-strand'. After the bacterium attaches to a plant cell, the T-strand and several types of Vir proteins are transferred to the plant through a transport channel. Inside the plant cell, the Vir proteins interact with the T-strand, forming a T-complex. This complex targets the nucleus, allowing the T-DNA to integrate into the plant genome and express the encoded genes.



Quantum optics

Cheat detection

Alice and Bob are getting divorced, but who keeps the dog? They could toss a coin for it, but as they now live apart, how can Bob be sure, if Alice tosses, that she won't lie about the result?

In most cryptographic problems of communication between two parties, the challenge is to eliminate third-party eavesdropping. But in the 'coin-tossing protocol', which is relevant to real-world situations such as remote signing of contracts, the issue is a lack of trust between the communicants. Can quantum

cryptographic methods stamp out cheating? Yes, they can, say G. Molina-Terriza *et al.* (*Phys. Rev. Lett.* **94**, 040501; 2005).

The authors have staged a quantum-optical enactment of a protocol that works as follows. The result is decided by many throws — the best of 100, say. Alice tosses each coin and encodes the outcome in the angular-momentum quantum state of a photon in an entangled pair. This encoding determines the quantum state of the other photon in the pair, which has been sent to

Bob. But Bob cannot determine the state of his photon until Alice sends him the information on her own photon — in essence, until she tells him how the coin fell. So he bets on the toss, and then Alice sends him a signal encoding the 'actual' outcome, which Bob can verify by performing a measurement on his own photon.

Because of its probabilistic nature, some such measurements will fail — Alice won't be able to encode the true outcome. But the crucial point is that if Alice is



systematically cheating, this shows up as a higher than average rate of 'failures', and Alice can do nothing to suppress this tell-tale signature of dishonesty.

Philip Ball

Agrobacterium-mediated genetic transformation has become a staple for basic plant research, as well as a principal means of generating transgenic plants for the agricultural biotechnology industry. However, at least for commercialization purposes, a tangled patent thicket complicates *Agrobacterium*-based plant-transformation technology. For this reason — as well as for more fundamental reasons of scientific interest — it would be interesting to determine whether other species of bacteria could, like *Agrobacterium*, serve as 'plant genetic engineers'. Although several previous papers have indicated that other bacterial species could incite plant tumours when engineered to carry a Ti-plasmid^{7–9}, molecular evidence for the actual transfer of DNA from these bacteria to plants has been lacking. Broothaerts *et al.*¹ now provide this evidence.

In their study, the authors used *Rhizobium* species NGR234, *Sinorhizobium meliloti* and *Mesorhizobium loti* — representatives of two different bacterial families. They first introduced a 'disarmed' Ti-plasmid that lacks a T-DNA region into these bacteria. This Ti-plasmid contained virulence genes, the protein products of which are necessary to act on a T-DNA and transfer it to plants (Fig. 2). Then Broothaerts *et al.* introduced a second plasmid containing a T-DNA region into these non-*Agrobacterium* species and into a disarmed *Agrobacterium* strain. This T-DNA contained a gene that encodes resistance to the antibiotic hygromycin, as well as *gusA*, a gene that directs the expression of the marker enzyme β -glucuronidase. To safeguard against misinterpretation of the data should the bacterial strains be mixed up, the T-DNAs introduced into *Agrobacterium* and non-*Agrobacterium* species contained different molecular 'tags'.

The authors used these bacteria to infect several different plant species, representing two major higher-plant groupings: species

from the dicots (which are highly susceptible to *Agrobacterium*-mediated transformation) included tobacco and the thale cress *Arabidopsis*; from the monocots (which are considerably less susceptible), the authors used rice. The generation of plants that express β -glucuronidase and are resistant to hygromycin indicated successful transformation by all bacterial species tested. This was confirmed by DNA-blot analysis and by the sequencing of T-DNA-plant-DNA junctions recovered from plants transformed with non-*Agrobacterium* species. Moreover, the junctions from these plants resembled those previously characterized as a result of *Agrobacterium*-mediated transformation. Therefore, the processing, transfer and integration of T-DNAs from the different bacterial species probably occurred by the same mechanisms.

Broothaerts and colleagues' results also suggest that non-*Agrobacterium* species are capable of the full range of genetic transformation mechanisms shown by their *Agrobacterium* counterparts. The transformation of *Arabidopsis* involved a 'flower dip' protocol, which targets female reproductive (germ-line) cells¹⁰. However, tobacco and rice were transformed via somatic tissues — leaves and callus cells, respectively — which probably involves a different mechanism¹¹.

The authors further found that although non-*Agrobacterium* bacterial species could be engineered to genetically transform several plant species, the transformation efficiency was relatively low, ranging from less than 1% to almost 40% of that of *Agrobacterium*-mediated transformation, depending on the species and transformation assays used. Broothaerts *et al.* note, however, that alterations in transformation conditions (the age and type of plant tissue inoculated, for instance) could enhance these frequencies. The low frequencies of *Agrobacterium*-mediated transformation of many 'recalcitrant' plant species have been

substantially increased through such manipulations over the past 20 years.

Broothaerts and colleagues' findings¹ will no doubt have implications for plant science and biotechnology. The ability of several bacterial species to transfer genes to plants also reinforces the possibility that such horizontal gene flow may have contributed to plant evolution. It is already known that the genomes of some plant species contain remnants of what were probably ancient transformation events by *A. rhizogenes*^{12,13}. In some instances, the *Agrobacterium*-derived transgenes are now expressed in a regulated manner in the host plant¹⁴, hinting that these genes have become an important part of the plant's genetic make-up. We may find that transgenes from other bacteria have likewise been assimilated by plants. Our understanding of the importance of horizontal gene flow in the evolution of higher organisms should be enhanced by knowledge of the mechanism of such transmission, and of the range of organisms that can participate in these events. ■

Stanton B. Gelvin is in the Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907-1392, USA.

e-mail: gelvin@bilbo.bio.purdue.edu

1. Broothaerts, W. *et al.* *Nature* **433**, 629–633 (2005).
2. DeClee, M. & DeLey, J. *Bot. Rev.* **42**, 389–466 (1976).
3. Gelvin, S. B. *Microbiol. Mol. Biol. Rev.* **67**, 16–37 (2003).
4. Bundock, P., den Dulk-Ras, A., Beijersbergen, A. & Hooykaas, P. J. J. *EMBO J.* **14**, 3206–3214 (1995).
5. de Groot, M. J. A., Bundock, P., Hooykaas, P. J. J. & Beijersbergen, A. G. M. *Nature Biotechnol.* **16**, 839–842 (1998).
6. Kunik, T. *et al.* *Proc. Natl Acad. Sci. USA* **98**, 1871–1876 (2001).
7. Klein, D. T. & Klein, R. M. *J. Bacteriol.* **66**, 220–228 (1953).
8. Hooykaas, P. J. J., Klapwijk, P. M., Nuti, M. P., Schilperoord, R. A. & Rorsch, A. J. *Gen. Microbiol.* **98**, 477–484 (1977).
9. van Veen, R. J. M. *et al.* *Mol. Plant-Microbe Interact.* **1**, 231–234 (1988).
10. Desfeux, C., Clough, S. J. & Bent, A. F. *Plant Physiol. Biochem.* **123**, 895–904 (2000).
11. Mysore, K. S., Kumar, C. T. R. & Gelvin, S. B. *Plant J.* **21**, 9–16 (2000).
12. Furner, I. J. *et al.* *Nature* **319**, 422–427 (1986).
13. Suzuki, K., Yamashita, I. & Tannaka, N. *Plant J.* **32**, 775–787 (2002).
14. Meyer, A. D., Ichikawa, T. & Meins, F. *Mol. Gen. Genet.* **249**, 265–273 (1995).

Cardiology

Solace for the broken-hearted?

Christine L. Mummery

The heart was thought to lack the capacity to regenerate after injury. But the identification of cells that can divide and mature into heart muscle suggests that the heart has repair mechanisms after all.

Heart attacks are fatal when they damage more than a quarter of the heart's left ventricle — killing off about 10^9 heart cells in the process. In patients who survive less severe attacks, dead heart cells are replaced by cells from the connective tissue called fibroblasts, which divide and migrate into the damaged area to form scar tissue. In these areas, the ventricular wall becomes thin and no longer contracts properly. Until recently, it seemed that the heart could not mend such an injury, but several reports now suggest that it might have some regenerative capacity after all¹⁻³. On page 647 of this issue, Laugwitz and colleagues⁴ provide compelling evidence that there are progenitor cells in the heart that might be able to engage in repair.

Like many specialized cells, fully developed heart cells do not divide and so are unable to patch up damage. What's needed are cells that are relatively unspecialized, so that they can divide, maintaining their own population and providing cells that can mature into heart cells. The cell population described by Laugwitz and colleagues seems to fit the bill.

These cells are distinguished by their expression of the *islet-1* gene (that is, they are $isl1^+$). By tracing the cells in a developing mouse embryo, Laugwitz *et al.*⁴ demonstrate

that the $isl1^+$ cells in the adult heart are remnants of a cardiac progenitor-cell population from the heart of the developing fetus. More excitingly, the authors also find $isl1^+$ cells in postnatal human hearts, and they can culture the mouse cardiac progenitors. Given an appropriate environment, the cultured cells divide and renew themselves, and they develop into what seem to be mature cardiac cells.

These cardiac progenitors are thus a potential source of cells for cardiac transplantation therapy. If they can be grown from human biopsies, then, in principle, tissue could be removed from a patient during surgery, and the cardiac progenitor cells could be amplified in culture and then implanted back into the patient's heart. The $isl1^+$ progenitor cells are found mainly in the heart atrium⁴, and taking atrial biopsies from patients has had no demonstrable negative effects. Alternatively, it may be possible to encourage the proliferation of cardiac progenitors *in situ* in the heart, and to enhance an endogenous repair mechanism that doesn't normally occur at a sufficient level to allow regeneration.

Cells derived from human embryos are being considered as another source for cardiac transplants⁵ because they can develop into almost any cell type. But use of such

embryonic stem cells is controversial, and using a patient's own cells would pose fewer rejection problems. Three recent studies⁶⁻⁸ have independently identified other primitive cells from the adult heart that are capable of dividing and developing into mature heart and vascular cells. These cardiac stem cells are distinct from the cardiac progenitors described by Laugwitz *et al.*⁴; both cell populations divide and renew themselves, but the progenitor cells are committed to becoming heart cells, whereas the stem cells have the potential to form several different cell types (Fig. 1).

Two of these studies found stem cells in the hearts of adult rats^{6,7}. These cells did not express *isl1*, but were isolated based on the presence of cell-surface proteins (either c-kit or Sca-1) that are usually associated with stem cells derived from bone marrow. The cardiac progenitors reported by Laugwitz *et al.* do not have these stem-cell markers, and so the cardiac progenitors and stem cells are molecularly distinct.

In the first study⁶, c-kit⁺ cells from rat hearts were shown to be self-renewing and capable of forming heart muscle cells and certain vascular cells. Although the heart muscle cells failed to contract spontaneously in culture, they did seem able to regenerate functional heart muscle when injected into a damaged heart. In the second study⁷, Sca-1⁺ c-kit⁻ cells from mouse hearts were able to develop into heart muscle when given intravenously after injury, although this was in part by fusing with the host heart cells.

Cardiac stem cells can also be grown from human biopsies as aggregates in suspension culture⁸. These so-called cardiospheres contain a mixture of cardiac cell types, and, when transplanted into mice after an acute heart attack, they formed vascular cells and heart muscle cells, albeit at a rather low frequency.

So, the race is on to find which cell type will be the most useful for heart repair. The outcome will depend on whether the cells can be produced in large numbers and induce long-term functional repair without causing irregular heart beats (arrhythmias). Fatal arrhythmias have occurred in patients who received their own skeletal muscle progenitor cells in an effort to repair heart damage, and this was attributed to poor communication between the transplanted cells and host heart muscle⁹. The risk of arrhythmia is minimal in mice because their fast heart rate (500 beats per minute) compensates for any disruption from transplanted cells. However, the risk could be significant in larger mammals, including humans, where heart rates are at least five-fold slower. Experiments in large animals such as pigs will be the best way to predict potential risks and assess the therapeutic value of cell transplantation into the heart before clinical trials are begun.

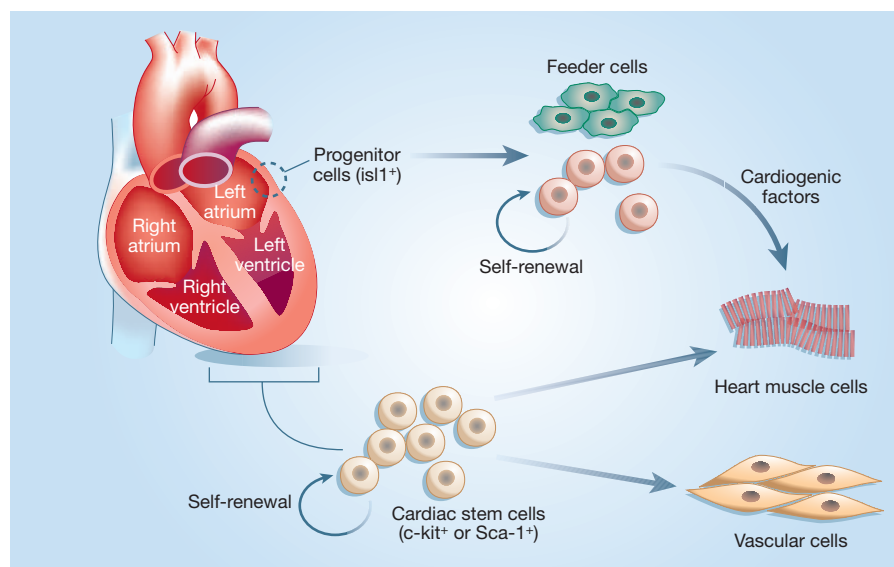


Figure 1 Cells for cardiac transplantation therapy. The cardiac progenitor cells described by Laugwitz *et al.*⁴ can renew themselves and develop into heart muscle cells; they are distinguished by the presence of the *islet-1* (*isl1*) protein. Cardiac stem cells⁶⁻⁸ can divide and develop into several cell types. These cells express proteins typical of stem cells from other organs (c-kit and Sca-1), and do not express *isl1*.

Adult bone-marrow cells were also thought to be a source of replacement heart muscle cells, but recently it was found that they cannot develop into cardiac cells^{10,11}. However, several small-scale clinical studies suggest that they may improve cardiac function when transplanted immediately after a heart attack. Although it is unclear how this might work, and most studies did not include a control patient group, no detrimental effects were observed in patients receiving their own bone marrow in the heart. For this reason, there have been calls to set up large-scale, controlled clinical trials^{12,13} using bone marrow, without extensive experiments in animals first.

Regenerating heart muscle by cardiac transplantation therapy is an ambitious goal, but with the current developments, it holds more than an abstract promise. ■

Christine L. Mummery is in the Hubrecht Laboratory and the Interuniversity Cardiology Institute of the Netherlands, University Medical Centre, Uppsalaalan 8, 3584 CT Utrecht, The Netherlands.

e-mail: christin@niob.knaw.nl

1. Nadal-Ginard B., Kajstura, J., Lerj, A. & Anversa, P. *Circ. Res.* **92**, 139–150 (2003).
2. Condorelli, G. et al. *Proc. Natl Acad. Sci. USA* **98**, 10733–10738 (2001).
3. Laflamme, M. A., Myerson, D., Saffitz, J. E. & Murry, C. E. *Circ. Res.* **90**, 634–640 (2002).
4. Laugwitz, K.-L. et al. *Nature* **433**, 647–653 (2005).
5. Kehat, I. et al. *Nature Biotechnol.* **22**, 1282–1289 (2004).
6. Beltrami, A. P. et al. *Cell* **114**, 763–776 (2003).
7. Oh, H. et al. *Proc. Natl Acad. Sci. USA* **100**, 12313–12318 (2003).
8. Messina, E. et al. *Circ. Res.* **95**, 911–921 (2004).
9. Menasche, P. et al. *J. Am. Coll. Cardiol.* **41**, 1078–1083 (2003).
10. Murry, C. E. et al. *Nature* **428**, 664–668 (2004).
11. Balsam, L. B. et al. *Nature* **428**, 668–673 (2004).
12. Strauer, B. E. et al. *Circulation* **106**, 1913–1918 (2002).
13. Wollert, K. C. et al. *Lancet* **364**, 141–148 (2004).

Climate change

Let all the voices be heard

D. M. Anderson and C. A. Woodhouse

It's a tough job to excavate trustworthy records about past temperatures from the palaeoclimate archives. The application of a fresh approach, in the form of wavelet analysis of the data, is a step forward.

Records of temperature during the past two millennia provide clues to the natural variation we might expect in the future. They also support attempts to partition recent warming into natural and anthropogenic components, and to measure the sensitivity of climate to greenhouse gases in the atmosphere. Such long records come only from natural archives, for example tree rings, ice cores and other 'proxy' evidence, and interpreting them has generated spirited debate. The crux of the issue is how much warmer or colder the average temperature has been in several century-long intervals of the past 2,000 years, the two most intensively studied intervals being the Medieval Warm Period (about AD 1200–1400) and the Little Ice Age (AD 1600–1850). Could we be in for one of these natural swings in the future? If so, how large would it be?

On page 613 of this issue¹, Moberg *et al.* present a multi-proxy reconstruction of annual temperature for the Northern Hemisphere for the past 2,000 years. Such a reconstruction is not new in itself, but it is unique in the way that the authors consider the information encoded in proxies of different temporal resolution. They use a clever statistical approach, based on 'wavelets'², to combine the climate information preserved in different proxies at different scales, and generate a reconstruction that reflects variability on scales ranging from the annual to the multi-centennial. In effect, the different scales of information in the proxies are

voices of different frequencies, and the aim is to hear them clearly.

Wavelet analysis is a powerful tool, already in use throughout science and engineering, and it is used here to extract, and then combine, the variance preserved in different proxies. The approach is both elegant and appropriate, providing a means of surmounting known limitations of the palaeoclimate archive. High-frequency (multi-year to multi-decade) variance is often blurred in all but the highest-fidelity ice or sediment records, and low-frequency (century-scale) variance is not usually well preserved in high-resolution tree-ring reconstructions, particularly those assembled from short overlapping series (the 'segment-length' curse³).

Using wavelets, the authors reconstruct a time series of Northern Hemisphere mean-temperature change from a suite of proxy records — hearing the high-frequency voices preserved in tree rings and the low-frequency voices retained in marine and lake sediments and other records. As a method of time-series analysis, wavelets offer several advantages, and are notably free from the assumption of stationarity (unchanging mean and variance) that makes most methods unsuitable for palaeoclimate time series.

The authors themselves recognize several shortcomings. The limited number of long tree-ring series available provides only a rough estimate of annual- to decadal-scale temperature variability in the Northern



100 YEARS AGO

The excavations [of Stonehenge] produced clear evidence touching the mode of erection... (1) The ground on the site it was to occupy was removed, the chalk rock being cut into in such a manner as to leave a ledge, on which the base of the stone was to rest, and a perpendicular face rising from it, against which as a buttress one side would bear when set up. From the bottom of this hole an inclined plane was cut to the surface, down which the monolith which had already been dressed was slid until its base rested on the ledge. (2) It was then gradually raised into a vertical position by means first of levers and afterwards of ropes. The levers would be long trunks of trees, to one end of which a number of ropes were attached... (3) As the stone was raised, it was packed up with logs of timber and probably also with blocks of stone placed beneath it. (4) After its upper end had reached a certain elevation, ropes were attached to it, and it was then hauled by numerous men into a vertical position, so that its back rested against the perpendicular face of the chalk which had been prepared for it.

From *Nature* 9 February 1905.

50 YEARS AGO

"Symposium on Genetics of Population Structure." Besides six papers, the Proceedings of the Symposium contain short comments by Dobzhansky, Lerner and Epling, the conclusions by Buzzati-Traverso and the scholarly but delightful address of thanks by Haldane... Scossiroli reported on the results of selection for bristle number in *Drosophila* populations after heavy X-ray irradiations... there occurred a spectacular response to selection in the 'high' direction but not in the 'low' one... It could be that X-ray-induced inheritable variation is mainly in the direction opposite to that for which natural selection had to work harder—a point well worth investigation at the threshold of the atomic age... The summing up... clearly showed how far the classic theoretical framework of population genetics has led... This change in outlook is essentially a shift of emphasis from the single gene to the integrated systems of chromosomes, the genotypes of the individuals and the whole gene pool of populations.

From *Nature* 12 February 1955.

Hemisphere; the impact of this shortcoming on the reconstruction has not been assessed. Only rudimentary spatial averaging was applied, considering neither the spatial representativeness of each proxy nor the difference in the quality of the records as temperature proxies. But these shortcomings do not compromise the main conclusions, and can be addressed in future work.

The results reveal substantially more variance at the century scale, and an amplitude of change equal to about 1 K (about the same as the borehole reconstructions⁴, and double the estimate based principally on tree rings⁵). Moberg and colleagues' reconstruction is consistent with results published earlier this year⁶ suggesting that tree-ring and other proxies produced with regression-based calibration methods commonly underestimate the amplitude of change.

What have we learned from the palaeo record of the past 2,000 years? Most tree-ring and high-resolution multi-proxy reconstructions show that Northern Hemisphere mean temperatures have varied about 0.5 K over the past millennium (temperatures were 0.5 K colder during the Little Ice Age relative to the 1961–90 mean). In contrast, reconstructions using boreholes⁴, and recent modelling analyses⁶, reveal a larger amplitude, close to 1.0 K. The results from Moberg *et al.* support the larger amplitude; even without the elegant mathematics, direct inspection of their compilation of low-frequency records indicates that the temperature change was about 1 K in many places.

Are the different proxies equally trustworthy and problem-free? Probably not. But the authors have taken a fruitful approach by attempting to combine information from different proxies in a way that uses the unbiased climate information found in each of them. Is it appropriate to consider change averaged over entire hemispheres? The emphasis of many palaeoclimatologists is shifting instead towards the identification of regional patterns of change. Many of the studies incorporated by Moberg *et al.* had that purpose — for example, even though the global climate system is interconnected, the tropical monsoon regions may have undergone different changes from those experienced by the high Arctic or the west Pacific.

The approach pioneered by Moberg *et al.* will serve equally well as a strategy to evaluate past climate change in each of these regions by combining information from different proxies. Indeed, the challenge for palaeoclimate researchers — and their funding agencies — is to produce multi-proxy reconstructions at the appropriate regional scale so that all of the voices can be heard. ■

D. M. Anderson and C. A. Woodhouse are in the Paleoclimatology Branch, National Climatic Data Center, Boulder, Colorado 80305–3328, USA.
e-mail: david.m.anderson@noaa.gov

1. Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M. & Karlén, W. *Nature* **433**, 613–617 (2005).
2. Torrence, C. & Compo, G. P. *Bull. Am. Meteorol. Soc.* **79**, 61–78 (1998).
3. Cook, E. R., Briffa, K. R., Meko, D. M., Graybill, D. A. & Funkhouser, G. *Holocene* **5**, 229–237 (1995).

4. Pollack, H. N. & Smerdon, J. E. *J. Geophys. Res.* **109**, doi:10.1029/2003JD004163 (2004).
5. Mann, M. E. & Jones, P. D. *Geophys. Res. Lett.* **30**, doi:10.1029/2003GL017814 (2003).
6. Von Storch, H. *et al. Science* **306**, 679–682 (2004).

Palaeoecology

Down to the woods yesterday

Peter D. Moore

What were European forests like following the last ice age and before the advent of agriculture? The pollen record in Ireland provides a unique perspective from which to examine ideas on the question.

The past may be a foreign country, and a pretty inaccessible one at that. But it does not deter people from exploring it. One of the objects of conservation is to retain or recreate primeval habitats, but this can only be achieved with sound knowledge of the structure and composition of past ecosystems, and the factors that sustained them.

Temperate deciduous forests (Fig. 1) are an example of such a habitat, and palaeoecologists are constantly probing the mists of time to answer some difficult questions — just how dense was the canopy of forests before the advent of agriculture, and what part did large grazing animals play in maintaining clearings and open conditions? This is a controversy that is of more than academic interest, and the latest foray, published by Fraser Mitchell in the *Journal of Ecology*¹, seeks to assure us that the forests were dense and that big grazers played a limited role in woodland dynamics.

The technique of pollen analysis, developed in the early twentieth century, revealed a progressive invasion of forest over western

Europe². In the mid-Holocene, between 8,500 and 5,500 radiocarbon years ago (BP), trees dominated the rain of pollen that fell on the lakes and bogs, where the pollen became preserved in stratified layers. Analysts interpreted this abundance of tree pollen as indicative of closed-canopy forest that had re-established itself following the last glacial retreat and then covered most of the landscape.

There were some puzzling features, however, such as the presence of 'non-arboreal pollen', the pollen of herbaceous plants, especially grasses and sedges, that persisted through the supposed closed-forest period. These open-habitat indicators became much more abundant as agricultural activities took hold and early farmers began to clear the forests around 5,500 BP. But the bogs and lakes where the pollen accumulated had their own local vegetation that included herbaceous plants, so their presence in the pollen record of the 'closed-forest' period was not regarded as surprising.

However, there are two additional aspects



Figure 1 Tree story in Europe. Pollen grains preserved in ancient lake sediments show that even before the advent of agriculture there were gaps in the forest. But were large grazing mammals responsible?

R. CUMMINS/CORBIS

to the data that require attention: how is it that species of oak maintain their populations in closed forest when their seedlings require light? And why is so much hazel pollen present when this understorey shrub flowers abundantly only in well-lit conditions? An alternative interpretation of the data was put forward by Vera³, who speculated that the mid-Holocene forest of western Europe was actually an open parkland, maintained in that condition by intensive grazing and browsing by large herbivorous mammals.

There is abundant archaeological evidence for the presence of many such herbivores in the European forests. Species of deer, elk (moose) and wild boar are still relatively common in many forest regions of Europe, and their impact on vegetation is often very apparent. In the mid-Holocene there were other large grazers present, such as wild horse, aurochs, species of bison, and beaver. But the experimental testing of any hypothesis relating to their role in forest dynamics is a problem, partly because we have no knowledge of their former abundance and density, and partly because their impact would probably have taken many centuries to be fully felt. Grazing experiments in the Netherlands served to encourage Vera in his views, but have not proved universally convincing. What is needed is a controlled experiment based on past conditions and operating over very long timescales. For this, one has to return to palaeoecology.

Mitchell¹ has come up with the proposal that the island of Ireland is a ready-made continental-scale grazing 'exclosure' experiment. Isolated from the British and continental mainland by rising sea-levels early in the Holocene, its vegetation has developed in the absence of many of the large grazers found farther east in Europe. Wild boar have been present, and red deer have been recorded from 4,000 BP onwards, although their early history in Ireland is unclear and they may well have been totally absent. All of the other large mammalian grazers were absent from Ireland throughout the Holocene.

The task facing Mitchell, therefore, was to compare forest development as documented by the pollen record in Ireland with that of continental Europe. To avoid the problems of local herb communities associated with lakes and bogs, he chose to study small, hollow sites in forests for his pollen data. These have the advantage that pollen arrives at them from a very restricted radius — a maximum of around 100 m (ref. 4) — so they provide a relatively uncontaminated record of local forest history. From the European Pollen Database, Mitchell selected 21 appropriate sites from Britain and continental Europe, and a further 15 from Ireland, and analysed them by multivariate statistical methods. He found no significant difference between forest

development on the continent, where large herbivores were active, and that in Ireland, where large grazers were virtually absent.

The approach is innovative in methodological terms, showing that the past is accessible for experimental analysis. But the outcome is also of great significance for forest managers and conservationists. Vera's proposals have raised questions about the appropriate target for forest management — should we be aiming for closed-canopy forest, or for open parkland (wood pasture) maintained by high densities of grazing stock? By rejecting Vera's hypothesis, Mitchell will bring those foresters deflected by the grazing proposal back to the closed-canopy model.

But not all questions have been answered.

Synthetic chemistry

Making a natural fuel cell

Marcetta York Darensbourg

The synthetic assembly of the active centre of hydrogen-producing enzymes adds to our understanding of their structure and function — and could produce new and useful materials that mimic these enzymes.

A host of microbes metabolize hydrogen with high efficiency by using enzymes known as hydrogenases¹. These enzymes are exquisite miniature hydrogen fuel cells, and are based on a combination of sulphur and iron atoms, and sometimes a single nickel atom². On page 610 of this issue, Tard *et al.* (Pickett and colleagues)³ demonstrate that a close analogue of the active site of all-iron hydrogenase can be synthesized in the laboratory, and that it can produce hydrogen in an electrochemical cell, given a supply of protons and electrons.

Preparation of an artificial analogue of the enigmatic H-cluster (the hydrogen-producing active centre of all-iron hydrogenase) has been a long-standing challenge for inorganic synthetic chemists. The problem is to reproduce the unusual spectroscopic signals that emanate from the 'black box' that is the unknown arrangement of the six iron atoms in the active centre. Early approaches made use of self-assembly reactions, which typically led to 4Fe4S cube structures, but sometimes resulted in larger, cage-like clusters^{4,5}. Although such expanded clusters may reasonably be expected to occur in nature, their spectroscopic signatures were not consistent with the active site of all-iron hydrogenase.

Some five years ago, protein crystallographers elucidated the structure of the functional centre of all-iron hydrogenase^{6,7}. The H-cluster is actually composed of two sub-sites: a normal cube-shaped 4Fe4S cluster is bridged by a sulphur atom from cysteine — an amino acid forming part of a surrounding

protein — to a second unit containing two iron atoms surrounded by ligands such as carbon monoxide and cyanide (Fig. 1a, overleaf). With this blueprint to hand, the challenge for chemists was to develop a synthesis method that maintains the integrity of the 4Fe4S and Fe–Fe precursors, and prevents them from rearranging into 6Fe6S expanded clusters.

The Fe–Fe organometallic portion of the H-cluster is a familiar one in chemistry. It can be modelled by a molecule that forms from iron and sulphur in the presence of carbon monoxide — $(\mu-S_2)Fe_2(CO)_6$, where μ denotes the bridging function of the sulphur atoms. This is a molecule of primordial origin, probably present in the reducing and harsh conditions of the ancient Earth, and also found today in oceanic thermal vents⁸.

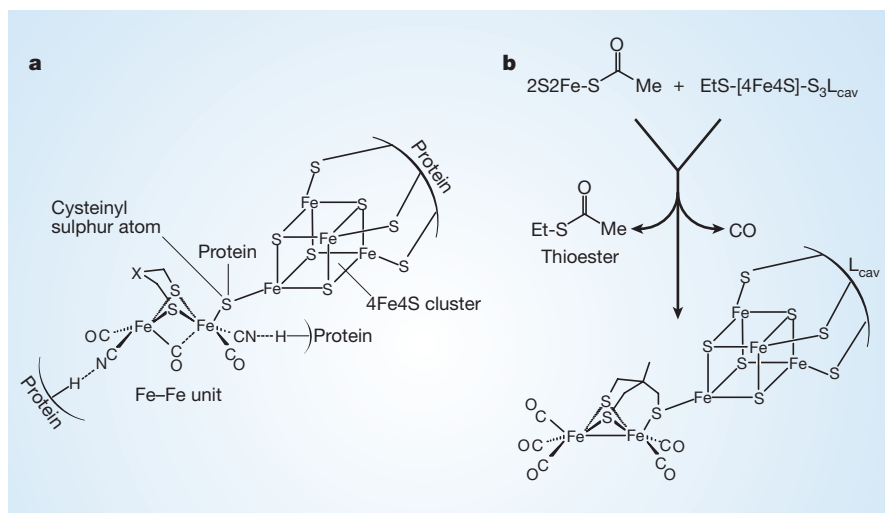
This compound acts as a template for condensation reactions involving small molecules, a process that possibly presaged the formation of biomolecules⁹.

Pickett and colleagues³ have developed a synthesis approach, in which an Fe–Fe organometallic unit is linked to a 4Fe4S cluster, by making use of sulphur-containing functional groups on independent subunits. These react and eliminate a small thioester molecule (Fig. 1b). As the synthetic precursor for the 4Fe4S cluster has four additional reactive sulphur-containing moieties, which are matched in nature by sulphur atoms in a protein's cysteine amino acids, it was necessary to isolate a single reaction site — hence the adoption of a large bowl-like, or cavitand, ligand^{4,5}. This ligand contains three

Peter D. Moore is in the Division of Life Sciences, King's College, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NH, UK.
e-mail: peter.moore@kcl.ac.uk

1. Mitchell, F. J. G. *J. Ecol.* **93**, 168–177 (2005).
2. Godwin, H. *History of the British Flora* (Cambridge Univ. Press, 1956).
3. Vera, F. W. M. *Grazing Ecology and Forest History* (CABI, Wallingford, 2000).
4. Sugita, S. *J. Ecol.* **82**, 881–897 (1994).

Figure 1 Synthesis in action. **a**, The active site, or 'H-cluster', of the all-iron hydrogenase, a hydrogen-metabolizing enzyme. It consists of a cube-shaped 4Fe4S cluster bridged to an Fe–Fe organometallic unit by a cysteinyl sulphur atom. **b**, Pickett and colleagues³ have made a close synthetic analogue of the H-cluster, using sulphur-containing precursors. The 4Fe4S and Fe–Fe units are linked up during a reaction in which a small thioester molecule and a carbon monoxide ligand are eliminated. L_{cav} is a large, bowl-shaped (cavitand) ligand^{4,5}, used to prevent expansion of the 4Fe4S cluster.



cysteinyllike sulphur donors in the cavity that bind to three of the cubane iron atoms, leaving only one sulphur donor that can react with the suitably modified Fe–Fe subsite¹⁰. And there was a chemical bonus as the elimination reaction occurred: a carbon monoxide ligand was simultaneously lost from the Fe–Fe subsite, allowing a cysteinyl-like sulphur bridge to form between two iron atoms, and thus further 'hard-wiring' the two subunits, precisely as in the H-cluster.

Pickett and colleagues³ show that the physical properties of the synthetic analogue are a good match for those of the enzyme, and that the Fe–Fe organometallic unit drains electron density from the 4Fe4S cluster. Most importantly, the chemical properties of the synthetic H-cluster indicate electrocatalysis of hydrogen production, consistent with the assumption that the Fe–Fe unit is a hydrogen-producing moiety. This process is the basis for the hydrogen-catalysing action of hydrogenase and is of great biotechnical interest: unlike today's commercial fuel cells, it does not require platinum — an expensive metal of limited availability — as the electrode. Can we imitate nature and construct platinum-free fuel cells? Pickett and colleagues have demonstrated a promising first step towards this goal.

But their work is not only of technological interest. Protein crystallography is a powerful technique — allowing chemists to recognize familiar patterns of atom–atom connections and overall geometry — but it produces only a snapshot of a factory in action. A close synthetic analogue of a biological molecule such as that reported by Pickett and colleagues, one that is structurally similar to its original, will allow researchers to assess the function of individual components that have evolved naturally to give efficiency and control. When chemists build an intricate molecule from known components through established pathways, it enhances their understanding of both its structure and its function. The

synthesis method described by Pickett and colleagues resolves the 'black box' spectral signals of the active centre of hydrogenase, and points towards a next generation of bio-inspired catalysts.

Marcetta York Darensbourg is in the Department of Chemistry, Texas A&M University, College Station, Texas 77843, USA.

e-mail: marcetta@mail.chem.tamu.edu

1. Vignais, P. M., Billoud, B. & Meyer, J. *FEMS Microbiol. Rev.* **25**, 455–501 (2001).
2. Cammack, R., Frey, M. & Robson, R. *Hydrogen as a Fuel*.

Learning from Nature (Taylor & Francis, London, 2001).

3. Tard, C. *et al. Nature* **433**, 610–613 (2005).
4. Beinert, H., Holm, R. H. & Munck, E. *Science* **277**, 653–659 (1997).
5. Averill, B. A., Herskovitz, T., Holm, R. H. & Ibers, J. A. *J. Am. Chem. Soc.* **95**, 3523–3534 (1973).
6. Peters, J. W., Lanzilotta, W. N., Lemon, B. J. & Seefeldt, L. C. *Science* **282**, 1853–1858 (1998).
7. Nicolet, Y. *et al. Struct. Fold. Des.* **7**, 13–23 (1999).
8. Cody, G. D. *et al. Science* **289**, 1337 (2000).
9. Adams, M. W. W. & Stiefel, E. I. *Curr. Opin. Chem. Biol.* **4**, 214–220 (2000).
10. Stack, T. D. P. & Holm, R. H. *J. Am. Chem. Soc.* **109**, 2546–2547 (1987).

Cancer

Catalyst of a catalyst

Stephen C. Kowalczykowski

Breast cancers arise when the BRCA2 protein is defective, but what does the normal enzyme do? Studies of a relative of BRCA2 reveal a capacity to initiate the repair of broken DNA by loading a repair protein.

The identification of the human *BRCA2* gene was a landmark discovery in the field of breast cancer¹. Mutations in this gene result in a dramatically increased predisposition to breast cancer, as well as to other tumour types. The *BRCA2* protein is known to be necessary for DNA repair, but its precise role has been unclear. Work from Pavletich and colleagues (Yang *et al.*², page 653 of this issue) now shows that a fungal counterpart of *BRCA2* can promote the loading of a key DNA-repair protein, Rad51, onto broken DNA. DNA breaks are a chronic problem for cells, arising from endogenous sources such as free radicals and the replication of damaged DNA, as well as from exogenous sources such as chemical mutagens and ionizing radiation. If not repaired correctly, these breaks cause chromosomal rearrangements that can result in cancer.

Initially, the sequence of the *BRCA2* gene¹ was both revealing and puzzling: the predicted protein was huge — 3,418 amino acids — but gave few hints to its function.

However, it is now known that the protein comprises eight consecutive repeats of an approximately 30-amino-acid sequence ('BRC repeats'), as well as a DNA-binding domain and a nuclear-localization signal. The first evidence that *BRCA2* might be involved in DNA repair came from the finding that it interacts physically with the repair protein Rad51, and that it is needed to form foci of Rad51 at sites of DNA breaks^{3,4}. The eight BRC repeats are particularly intriguing in this regard: they interact with Rad51, both *in vivo* and *in vitro*^{3–6}, but in isolation they have the puzzling property of inhibiting Rad51 activity^{4,5}. Thus, although it was thought that *BRCA2* would be involved in Rad51-induced DNA repair, its precise role remained enigmatic.

Nonetheless, the discovery of the interaction with Rad51 did provide an important clue to exactly how *BRCA2* might function. Rad51 is essential for the repair of DNA breaks by homologous recombination. In this process, Rad51 needs to assemble into a

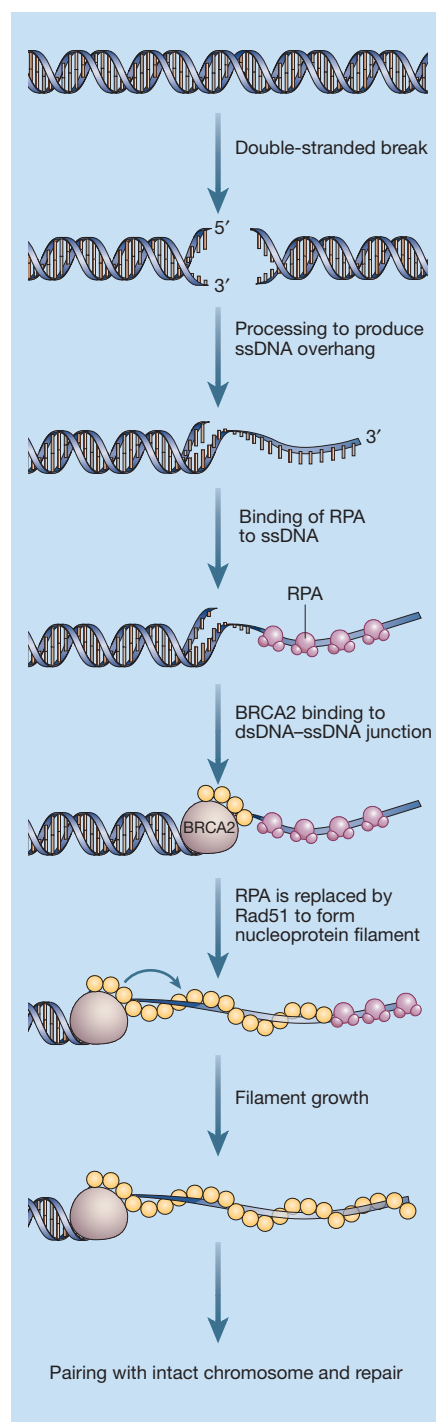


Figure 1 Model for the function of BRCA2, based on the findings of Yang *et al.*². When a chromosome breaks, its DNA must be repaired. The first step in one repair process, homologous recombination, involves the processing of the DNA break to produce a single-stranded (ss) region with a 3' overhang. Replication protein A (RPA) binds to the single-stranded region. Yang *et al.* propose that BRCA2 then binds to the junction between the double-stranded (ds) and single-stranded regions, and recruits the Rad51 protein, which ousts the RPA, forming a Rad51 'nucleoprotein filament'. Pairing with the matching DNA region on the intact chromosome copy then enables the break to be repaired by homologous recombination.

helical structure by binding to the single-stranded DNA (ssDNA) that is formed after one DNA strand is degraded (processed) at the site of a break. The resulting 'nucleoprotein filament' finds matching (homologous) DNA in the intact copy of the damaged chromosome, and then exchanges DNA strands to provide a template that is used to replicate the damaged or lost DNA. Sealing the DNA nicks completes the repair. Assembly of the Rad51 nucleoprotein filament is inhibited by replication protein A (RPA)^{7,8}, which binds to ssDNA, blocking Rad51 binding. To accelerate filament assembly, proteins known as mediators intervene^{9,10}, recruiting Rad51 to the RPA-DNA complex and replacing RPA by Rad51. Might BRCA2 also function as a mediator?

That notion was given a physical basis from two separate crystal structures. One structure¹¹ revealed that BRCA2 possesses a DNA-binding domain with separate regions for binding ssDNA and double-stranded DNA (dsDNA). The second⁶ showed that a BRC repeat binds to the core of Rad51, and 'mimics' the structure of an adjacent Rad51 monomer. Together, these structures suggested that the intact BRCA2 protein might bind both DNA and Rad51, and help deliver Rad51 to processed DNA breaks¹².

Yang *et al.*² have now defined a function for BRCA2 by studying a simpler relative of the human protein. *BRCA2* genes are not unique to mammals, and homologues exist in plants and fungi¹³. The homologue from the fungus *Ustilago maydis* is *BRH2* (for 'BRCA2 homologue'), and it encodes a small (1,075-amino-acid) and relatively simple protein, comprising just one BRC repeat and a DNA-binding domain similar to that of BRCA2 (ref. 13). Cells defective in *BRH2* are sensitive to ultraviolet and ionizing radiation, are defective in the recombinational repair of DNA damage and, consequently, undergo chromosomal rearrangements — features common to cells with mutated *BRCA2*.

Yang *et al.* show convincingly that the Brh2 protein can nucleate the assembly of Rad51 onto ssDNA that is coated with RPA, efficiently producing a Rad51 nucleoprotein filament that can promote DNA strand exchange. Furthermore, even though Brh2 can promote the formation of Rad51 filaments on ssDNA, it acts preferentially at a dsDNA-ssDNA junction that has a particular type of single-stranded overhang (a 3'-terminated overhang). Brh2 binds to this junction to initiate the replacement of RPA by Rad51 in the 5' to 3' direction along the ssDNA (Fig. 1). Other recombination mediator proteins act by one of two broad mechanisms: there are those that act in stoichiometric complexes with RPA and Rad51, and those that act sub-stoichiometrically (as catalysts) relative to the amount of RPA and Rad51 (refs 7–10,14). Brh2 belongs to this second class. Thus, Brh2 is confirmed

as a mediator of Rad51 assembly — a catalyst of a catalyst in DNA repair.

What is remarkable about these results is that Brh2 loads Rad51 onto ssDNA at the junction with dsDNA without any other factors, energy source or phosphorylation/de-phosphorylation cycle. This activity would be unique were it not for the behaviour of one other set of proteins that functions in recombination — the bacterial RecFOR protein complex¹⁴. RecFOR loads the bacterial homologue of Rad51 (RecA) onto complexes of ssDNA and ssDNA-binding protein, specifically at the junction with dsDNA, and with the same polarity. Moreover, only catalytic amounts of RecFOR are required. Thus, Brh2 is a precise functional homologue of the RecFOR complex.

Yang and colleagues' work³ on Brh2 has clear implications for our understanding of BRCA2 function. First, it makes it virtually certain that BRCA2 is a catalyst of human Rad51 assembly at the sites of processed DNA breaks. This explains why Rad51 foci do not form in cells that lack functional BRCA2. Second, the inhibitory effects of BRC peptides can now be understood: such peptides can bind to Rad51 but, lacking the DNA-binding domain, cannot deliver Rad51 to the ssDNA, and therefore behave as inhibitors. And finally, the DNA-binding domain is required to bring BRCA2 — presumably with Rad51 bound to the BRC repeats — to processed DNA breaks. Once there, Rad51 can be transferred onto the ssDNA to form a filament; or, as Yang *et al.* propose, a BRC-bound Rad51 can initiate the formation of a filament without dissociating from BRCA2. Nucleation of Rad51 assembly on the ssDNA then initiates DNA repair. Further clarification of the mechanism of BRCA2 function should improve our understanding of the processes involved in tumour development and, hopefully, of how to correct the underlying defects. ■

Stephen C. Kowalczykowski is in the Sections of Microbiology and of Molecular and Cellular Biology, Center for Genetics and Development, University of California, Davis, 1 Shields Avenue, Davis, California 95616-8665, USA.
e-mail: skowalczykowski@ucdavis.edu

1. Wooster, R. *et al.* *Nature* **378**, 789–792 (1995).
2. Yang, H., Li, Q., Fan, J., Holloman, W. K. & Pavletich, N. P. *Nature* **433**, 653–657 (2005).
3. Wong, A. K. C., Pero, R., Ormonde, P. A., Tavtigian, S. V. & Bartel, P. L. *J. Biol. Chem.* **272**, 31941–31944 (1997).
4. Chen, C. F. *et al.* *J. Biol. Chem.* **274**, 32931–32935 (1999).
5. Davies, A. A. *et al.* *Mol. Cell* **7**, 273–282 (2001).
6. Pellegrini, L. *et al.* *Nature* **420**, 287–293 (2002).
7. Shinohara, A. & Ogawa, T. *Nature* **391**, 404–407 (1998).
8. New, J. H., Sugiyama, T., Zaitseva, E. & Kowalczykowski, S. C. *Nature* **391**, 407–410 (1998).
9. Beernink, H. T. & Morrical, S. W. *Trends Biochem. Sci.* **24**, 385–389 (1999).
10. Sung, P., Krejci, L., Van Komen, S. & Sehorn, M. G. *J. Biol. Chem.* **278**, 42729–42732 (2003).
11. Yang, H. *et al.* *Science* **297**, 1837–1848 (2002).
12. Kowalczykowski, S. C. *Nature Struct. Biol.* **9**, 897–899 (2002).
13. Kojic, M. *et al.* *Mol. Cell* **10**, 683–691 (2002).
14. Morimatsu, K. & Kowalczykowski, S. C. *Mol. Cell* **11**, 1337–1347 (2003).

Obituary

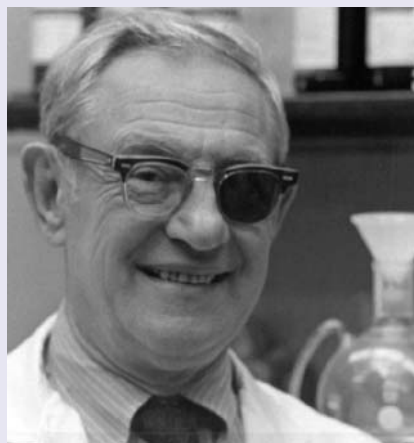
Julius Axelrod (1912–2004)

Julius Axelrod, known to friends and colleagues affectionately as 'Julie', died on 29 December 2004, following a career that was highlighted by so many extraordinary discoveries about how drugs act that his place among the twentieth century's greatest biochemical pharmacologists is secure. The research for which he shared the Nobel Prize in Physiology or Medicine in 1970 was only a small, and perhaps not even the most important, part of his opus.

Axelrod's life story is an inspiration to those who care about science but lack elaborate academic credentials. The son of a poor basket-maker in New York City, he graduated from the City College of New York and applied to numerous medical schools but was rejected by all of them. For 11 years he worked as a technician in a New York City health laboratory, where he was assigned the task of monitoring the vitamin content of foods. Here he honed his skills in developing simple, sensitive and specific methods for measuring drugs.

A turning point was his department's assignment to clarify why phenacetin and acetanilide — the principal ingredients, along with aspirin, in major headache remedies of the day — led to dangerously high levels of methaemoglobin, a product of haemoglobin metabolism, in patients' blood. In 1946, he sought the assistance of Bernard Brodie, then at the Goldwater Memorial Hospital, thus commencing a career-transforming collaboration. With Brodie he showed that acetanilide and phenacetin are respectively metabolized to aniline and *p*-phenetidin — the toxic culprits. Another metabolite, *N*-acetyl-*p*-aminophenol or acetaminophen (paracetamol), is probably responsible for the analgesic effects. This discovery, never patented, made up Axelrod's first two publications, and led to the marketing of this metabolite as Tylenol, the world's most widely used analgesic.

In 1950, Axelrod moved to the newly formed National Institutes of Health (NIH) in Bethesda. He continued working in Brodie's laboratory but with increasing independence, elucidating the metabolism of caffeine, ephedrine and amphetamine in animals. Although he had never worked with enzymes, he wanted to understand in greater detail how drugs are metabolized, and so incubated various subcellular fractions with amphetamine and ephedrine. Using membrane fractions from the liver, he discovered a previously unknown group of enzymes that require NADPH (the reduced form of nicotinamide adenine dinucleotide phosphate) and



From pharmacologist to neuroscientist

molecular oxygen. This is the cytochrome P450 family of drug-metabolizing enzymes, and their discovery represents a classic advance in pharmacology.

All of these discoveries were made while Axelrod was only a technician. Responding to pressure from friends, he took a year's leave of absence, enrolled at George Washington University and, at the age of 42, received his PhD. With a doctorate in hand, Axelrod was made head of pharmacology in the laboratory of clinical science at the National Institute of Mental Health.

Seymour Kety, the laboratory director, was intrigued by reports that implicated the abnormal metabolism of adrenaline (also known as epinephrine, a neurotransmitter) in schizophrenia. To understand how adrenaline is normally metabolized, Axelrod exploited a brief abstract reporting that adrenaline-producing tumours metabolize adrenaline to a product that is methylated on one of the two hydroxyl groups of its catechol ring. Axelrod soon discovered an enzyme that uses *S*-adenosylmethionine to methylate catechols; he dubbed it catechol-*O*-methyltransferase. Using radiolabelled *S*-adenosylmethionine, Axelrod went on to identify several other methyltransferases. These include the adrenal enzyme that generates adrenaline from noradrenaline, and one that forms melatonin, a substance unique to the pineal gland.

Kety had set up research wards to compare the metabolism of radiolabelled adrenaline in schizophrenics and normal controls, and provided Axelrod with samples of [^3H]adrenaline for metabolic studies. When [^3H]adrenaline or

[^3H]noradrenaline was injected into rats and cats, the radiolabel concentrated in tissues enriched in sympathetic-nerve endings. But lesions of the sympathetic nerves abolished the accumulation. So Axelrod postulated that noradrenaline, the neurotransmitter used by sympathetic nerves, is inactivated by a 're-uptake' system that involves the transmitter being pumped back into the nerve ending that had released it. Adrenaline, released from the adrenal gland, is also taken up into sympathetic nerves.

Neurotransmitter inactivation had previously been thought to be primarily enzymatic, because the first known transmitter, acetylcholine, is inactivated by an enzyme. But it soon became evident that re-uptake inactivation is the rule rather than the exception for most neurotransmitters. Axelrod also found major drugs that inhibit re-uptake, and his discovery that antidepressants work in this way spawned the modern generation of antidepressants, exemplified by fluoxetine (Prozac). For this research he shared the Nobel prize with Ulf van Euler, who established that noradrenaline is a neurotransmitter, and Bernard Katz, who showed that neurotransmitters are stored in synaptic vesicles in nerve endings.

Axelrod's gifts as a creative scientist were matched by his gifts as a mentor. His extraordinary discoveries emerged from a tiny laboratory of only three or four scientists. Although he carefully structured concrete projects for new students, he soon weaned them, encouraging independent thinking by the "unconditional positive regard" (to quote psychologist Carl Rogers) that he lavished on every student's efforts. He maintained an active laboratory at the NIH until a few years ago, and subsequently visited his small office there regularly. His wife Sally died in 1992; he is survived by his sons Paul and Fred and three grandchildren.

Like almost all of Axelrod's trainees, my entire scientific career was predicated upon the lessons learned in my two years (1963–65) with him. Many of us have tried to divine Julie's gift for recognizing important questions that can be addressed by simple means, and for devising experiments so elegantly efficient that, with just 25 test tubes, he could answer four or five important scientific questions. We all have our own ideas about Julie's magic formula, but I suspect he has taken the real answer with him to the grave.

Solomon H. Snyder

Solomon H. Snyder is in the Department of Neuroscience, Johns Hopkins Medical School, 725 North Wolfe Street, Baltimore, Maryland 21205, USA.
e-mail: ssnyder@bs.jhmi.edu

Immunology

A nod in the right direction

Science **307**, 731–734; 734–738 (2005)

A gene called *Nod2* is often mutated in people with Crohn's disease, in which chronic inflammation in the small intestine causes pain and diarrhoea. Two groups have begun to unravel how the altered gene might increase susceptibility to the condition.

Koichi S. Kobayashi *et al.* show that mice genetically engineered to lack *Nod2* protein are more susceptible to bacterial infection, and that immune cells lacking *Nod2* no longer recognize a bacterial coat protein or trigger the normal protective immune response. The authors also find that *Nod2* switches on the production of a group of antimicrobial peptides in the intestine that are important in suppressing infection by pathogenic bacteria.

Meanwhile, Shin Maeda *et al.* find that mice engineered to carry human-type mutations in *Nod2* are also more susceptible to bacterially induced intestinal inflammation. They suggest that the protein controls the activity of two molecules involved in inflammation, NF- κ B and interleukin-1 β .

The results suggest that *Nod2* mutations might promote Crohn's disease by hobbling normal immune responses, rather than by triggering the disease itself — a finding that may point to new ways of treating the condition.

Helen Pearson

Metabolism

From fats to fat

Cell **120**, 261–273 (2005)

We all know that a high-fat diet leads to increased body fat and clogged arteries, but the metabolic details of this cause-and-effect chain are still not completely clear. A new link has now been added, with the discovery that high-fat foods stimulate the liver's production of PGC-1 β , a transcriptional coactivator protein that promotes the creation of fatty molecules such as cholesterol and triglycerides.

Mice fed a diet rich in saturated fats but low in cholesterol showed increased PGC-1 β expression in the liver, report Jiandie Lin and colleagues. This in turn binds to and coactivates the sterol responsive element binding protein (SREBP) family of transcription factors, which can boost the activity of fat-creating genes.

The finding adds to the picture of how a fatty diet leads to increased blood levels of cholesterol and triglycerides, which can cause heart disease and atherosclerosis. Controlling the expression of PGC-1 β in the liver might help to minimize the health impacts of a high-fat diet, the



Environment

Congestion success

Atmos. Environ. **39**, 1–5 (2005)

The London congestion charging scheme has been causing controversy since it began in February 2003. Some businesses claim that the charge of £5 (US\$9) to drive in central London deters shoppers, but the scheme has succeeded in reducing traffic jams.

Unusual weather conditions have made it difficult to assess the full impact of the scheme on air pollution, but using a combination of traffic-monitoring data and emissions modelling, Sean D. Beevers and David C. Carslaw obtained some interesting findings. Between 2002 and

2003, the emission of nitrogen oxide compounds was reduced by about 12% inside the charging zone, while levels increased by roughly 1.5% on main roads just outside the zone.

Emission of small particles decreased by about 12% inside the zone, and by 1.5% just outside. Carbon dioxide emissions in the zone are down by almost 20%.

A significant factor in these reductions is that cars produce lower emissions when they travel faster: average speeds in the charging zone rose from 19 to 23 km h⁻¹ in 2003.

Although the authors admit a large uncertainty in the calculations, they say that their results are likely to underestimate the benefits of the scheme.

Mark Peplow

authors suggest — although a means to the same end is to eat less saturated fat in the first place.

Michael Hopkin

Animal behaviour

The sound of sustenance

Anim. Behav. doi:10.1016/j.anbehav.2004.06.020 (2005)

Bottlenose dolphins apparently use their sophisticated echolocation powers only sparingly in the wild, and the prevalence in their diet of fish that produce sound in various ways led to the idea that they listen passively for food.

Damon P. Gannon *et al.* provide evidence to support this idea. The researchers conducted playback experiments off the coast of Florida to determine the foraging behaviour of bottlenose dolphins. Gannon *et al.* played sound recordings of several species of prey fish in the water, as well as the sounds of the snapping shrimp — which the dolphins are not known to eat — as a control.

When the dolphins heard the fish sounds, they immediately changed their direction of travel, turning to pursue the fish and upping the previously low rate of echolocation. Control noises, on the other hand, elicited no change in behaviour. This dominance of passive listening in the initial stages of hunting suggests that echolocation incurs a heavy energetic or ecological cost, the authors say.

Roxanne Khamsi

Oceanography

Global warming all at sea

Geophys. Res. Lett. **32**, L02604 (2005)

Discussions of global warming tend to focus on the mean temperature of the atmosphere at the Earth's surface, which has warmed by about 0.8 °C over the past century. But that tells only part of the story, according to S. Levitus and colleagues. They have brought together more than four decades of measurements to calculate that the world ocean has, on average, warmed by 0.037 °C since 1955.

Sounds paltry? It certainly isn't. Water has a very large heat capacity, which means that it takes a lot of heat to raise its temperature. This apparently small increase in ocean temperature has been brought about by an enormous input of heat, accounting for 84% of the total increase in heat content of the Earth system. To put this in perspective, an ocean warming of 0.1 °C would, if all the heat were discharged to the atmosphere, warm the planet's surface by 100 °C.

A large part of the warming has occurred in the upper 700 metres of the oceans, although the distribution of heat varies geographically: the Atlantic Ocean has taken up more heat than the Pacific and Indian Oceans combined.

Philip Ball

Motor control of flexible octopus arms

The octopus borrows a jointed-vertebrate strategy to transfer an item between points.

Animals with rigid skeletons can rely on several mechanisms to simplify motor control^{1–9} — for example, they have skeletal joints that reduce the number of variables and degrees of freedom that need to be controlled. Here we show that when the octopus uses one of its long and highly flexible arms to transfer an object from one place to another, it employs a vertebrate-like strategy, temporarily reconfiguring its arm into a stiffened, articulated, quasi-jointed structure. This indicates that an articulated limb may provide an optimal solution for achieving precise, point-to-point movements.

Octopuses use their arms to reach a target by generating a stereotypical movement in which only three degrees of freedom are controlled^{10–12}. Once the object has been grasped at some point along the arm, the animal needs to execute a point-to-point movement to transfer this object to the mouth. This action of bringing the end-effector (the site on the arm where the object is grasped) to a specific target (the mouth) resembles the execution of motor tasks by articulated arms.

Fetching movements by octopuses (*Octopus vulgaris*) that are allowed to behave freely are most frequently used to bring food towards the mouth (Fig. 1a; for movie, see supplementary information). These are stereotypical movements that were recognizable in roughly 100 filmed trials, and they can be initiated by placing a piece of food anywhere along the arm (for methods, see supplementary information).

The animal's suckers rapidly grasp the food item and a distal and a medial bend form along the arm shortly afterwards; the base of the arm serves as a third (proximal) bend. The locations of the object and of these three bends together define a quasi-articulated structure (Fig. 1a, b), which then rotates mainly about the medial bend in order to bring the distal bend to the base of the arm. Finally, rotation about the distal bend brings the object to the mouth.

Kinematic analysis of the fetching movements (see supplementary information) reveals that the location of the object and the two bends divide the quasi-articulated structure into three segments, each of length L : proximal (L_1), medial (L_2) and distal (L_3) (Fig. 1b). To test whether the bends stay at fixed locations along the arm, we measured each segment length during the execution of the different movements. The slopes of plots of L_1 and L_2 against time did not differ significantly from zero (-0.05 ± 0.14 , $P=0.1$; 0.04 ± 0.20 , $P=0.07$, respectively, in one-sample t -test; data not shown), so L_1 and L_2 must have remained nearly constant. However, L_3

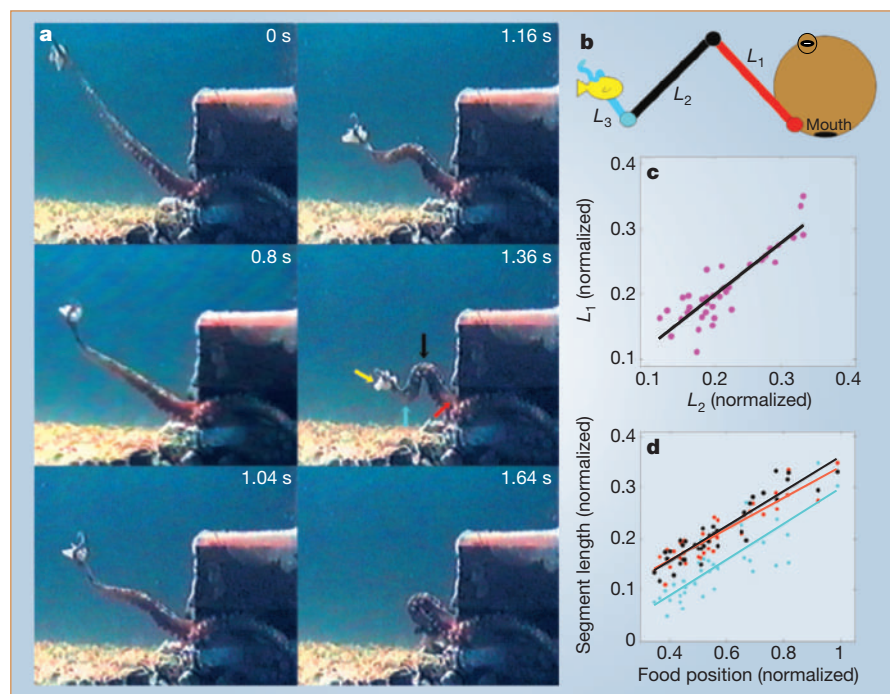


Figure 1 Octopus fetching movements involve a quasi-articulated structure that is based on three bending sites, or 'joints'. **a**, Sequence of video images taken during a fetching movement. Yellow arrow, food item; blue, black and red arrows indicate distal, medial and proximal 'joints', respectively. **b**, Quasi-articulated structure, showing arm-segment nomenclature and coloured as in **a**. **c**, Correlation between the length of the medial and proximal segments ($n=36$); each point represents one trial. **d**, Correlation between the position of the food object along the arm and the different segment lengths (coloured as in **a**, **b**) ($n=36$). Segment lengths and the position of the food item are normalized with respect to the total length of the arm.

showed a tendency to shorten, with the slope of L_3 against time differing significantly from zero (-0.35 ± 0.26 , $P=0.03$; one-sample t -test). These results indicate that the bends are behaving like the joints in a skeletal structure.

The quasi-articulated structure shows a strikingly invariant geometry for all of the recorded movements (Fig. 1c): L_1 and L_2 are positively correlated (F -test, $R^2=0.78$, $P=1.6 \times 10^{-12}$) with a slope of 0.8 ± 0.15 (mean $\pm$ confidence interval). Therefore, during each fetching movement the proximal (L_1) and medial (L_2) segments are almost equal in length, a geometry that resembles the general structure of vertebrate arms and some arthropod appendages.

What is unique to the octopus, and contrasts with the appendages of rigid skeletal animals, is that the segment lengths are positively and linearly correlated with the site on the arm at which the food is grasped (Fig. 1d, and see supplementary information). This result indicates that the configuration of the quasi-articulated structure is dynamically adjusted for each movement according to the position of the grasping site.

Electromyogram recordings made during the fetching process revealed that the quasi-articulated structure is actively contracted,

or stiffened, and that the degree of this muscle activity correlates best with the segment lengths (see supplementary information). This finding probably reflects an underlying control strategy for stabilizing the structure against water-drag forces.

It is surprising, given the large number of possible ways in which a flexible arm could convey an object to the mouth, that the octopus uses a quasi-articulated structure that resembles the multijointed, articulated limbs of animals with rigid skeletons. Fetching seems to be an example of evolutionary selection of solutions that are similar even though they are based on quite different mechanisms—on morphology in arthropod and vertebrate limbs, and on stereotypical motor control in the octopus. This functional convergence suggests that a kinematically constrained, articulated limb with two segments of almost equal length is the optimal design for accurately moving an object from one point to another.

Germán Sumbre*†, Graziano Fiorito‡, Tamar Flash§, Binyamin Hochner*

*Department of Neurobiology and Interdisciplinary Center for Neuronal Computation, Institute of Life Sciences, Hebrew University, Jerusalem 91904, Israel
e-mail: bennyh@lobster.lsh.huji.ac.il

‡Laboratorio di Neurobiologia, Stazione Zoologica di Napoli "A. Dohrn", Naples 80121, Italy

§Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel

†Present address: Division of Neurobiology, Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA

- Flash, T. & Hogan, N. *J. Neurosci.* **5**, 1688–1703 (1985).
- Harris, C. M. & Wolpert, D. *Nature* **394**, 780–784 (1998).
- Feldman, G. *Biophysics* **11**, 766–775 (1966).
- Bizzi, E., Hogan, N., Mussa-Ivaldi, F. A. & Giszter, S. *Behav. Brain. Sci.* **15**, 603–613 (1992).
- Bizzi, E., Tresch, M. C., Saltiel, P. & d'Avella, P. *Nature Rev. Neurosci.* **1**, 101–108 (2000).
- Soechting, J. F., Buneo, C. A., Herrmann, U. & Flanders, M. *J. Neurosci.* **15**, 6271–6280 (1995).
- Flash, T. & Sejnowski, T. J. *Curr. Opin. Neurobiol.* **11**, 655–662 (2001).
- Sabes, P. N. *Curr. Opin. Neurobiol.* **10**, 740–746 (2000).
- Wolpert, D. M. & Ghahramani, Z. *Nature Neurosci. suppl.* **3**, 1212–1217 (2000).
- Gutfreund, Y. et al. *J. Neurosci.* **16**, 7297–7307 (1996).
- Gutfreund, Y., Flash, T., Fiorito, G. & Hochner, B. *J. Neurosci.* **18**, 5976–5987 (1998).
- Sumbre, G. et al. *Science* **293**, 1845–1848 (2001).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Laser technology

Source of coherent kiloelectronvolt X-rays

Generating X-rays that have the properties of laser light has been a long-standing goal for experimental science. Here we describe the emission of highly collimated, spatially coherent X-rays, at a wavelength of about 1 nanometre and at photon energies extending to 1.3 kiloelectronvolts, from atoms that have been ionized by a 5-femtosecond laser pulse. This means that a laboratory source of laser-like, kiloelectronvolt X-rays, which will operate on timescales relevant to many chemical, biological and materials problems, is now within reach.

High-order harmonic generation in an atomic gas ionized by a femtosecond laser pulse has been successful in generating coherent extreme-ultraviolet radiation^{1,2}. The coherent laser field forces the atoms to radiate waves that oscillate in the same way in a macroscopic volume, resulting in the coherent build-up of extreme-ultraviolet radiation in a collimated laser-like beam. The freed electrons cause the generating laser wave and emerging short-wavelength radiation to travel at different speeds and hence limit the useful interaction length, over which the atomic dipoles radiate in phase. Suppression of this de-phasing effect by suddenly turning on the laser field³, or by modulating the propagation constant in the generation medium⁴, results in an extension of harmonic emission to several hundred electronvolts.

We advanced the former approach³ by pushing ultrafast laser technology to extend the spectral range of coherent radiation to

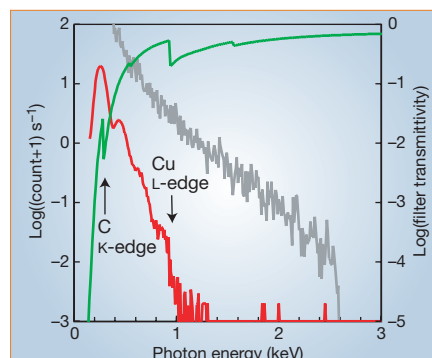


Figure 1 Photon energy spectrum of coherent soft-X-ray emission from a helium gas (red line), measured using a multichannel analyser (Oxford) and a silicon–lithium semiconductor detector cooled by liquid nitrogen. The soft X-ray beam is filtered by propagation through 100-cm helium (3 millibar), 100-nm copper and 100-nm aluminium filters and a 300-nm AP1.3 window (including carbon). Green line, overall transmittivity of these filters; grey line, calculated spectrum of radiation emitted by individual helium atoms exposed to 5-fs pulses with a peak intensity of $1.4 \times 10^{16} \text{ W cm}^{-2}$.

the kiloelectronvolt regime. A multistage titanium:sapphire laser⁵ delivers near-infrared pulses of under 15 fs, which are spectrally broadened in a neon-gas-filled capillary and compressed by chirped mirrors in order to obtain 720-nm, 5-fs, 0.2-terawatt pulses at a repetition rate of 1 kHz. The linearly polarized pulses propagate through a 0.4-mm-thick jet of helium atoms (pressure, 0.5 bar; $1/e^2$ diameter of laser beam, $w_0 = 30 \mu\text{m}$). The short-wavelength radiation emerges from the gas jet in a laser-like beam.

We analysed the beam's spectral distribution using energy-dispersive and wavelength-dispersive X-ray spectrometers, which separate different spectral components according to their photon energy or wavelength, respectively. The combination of aluminium, carbon and copper filters with the helium background atmosphere in the vacuum system blocks light that has photon energies below 200 eV before detection (for methods, see supplementary information). From a knife-edge scan across the X-ray beam, we evaluated a beam-divergence angle of 0.2 milliradians for the spectral range above 200 eV. This divergence implies perfect coherence of the atomic dipoles within a macroscopic volume of diameter of $\geq 13 \mu\text{m}$ and $\geq 4 \mu\text{m}$ at photon energies of 0.3 keV and 1 keV, respectively. The emitted X-ray beam seems to be diffraction-limited to within a factor of five (see supplementary information).

Figure 1 shows the spectrum of the generated radiation, as transmitted through the filters and recorded with the energy-dispersive X-ray spectrograph. The appearance of the copper L-edge at about 950 eV is evidence for harmonic radiation at 1 keV. X-rays were detected beyond 1.3 keV (see supplementary information). The key to this progress is the temporal intensity gradient in the driving pulse, which allows some 25% of the helium atoms to be ionized within half a cycle before

the pulse peak⁶. The electrons detached within this time are pushed in the most intense half-cycle back to the atomic core. Under these conditions, our simulations^{6,7} predict harmonic emission from individual helium atoms up to 2.5 keV (Fig. 1).

We have shown that atoms ionized within a few femtoseconds emit coherent, kiloelectronvolt soft X-rays. As the few-cycle pulse can propagate free of substantial distortion over distances that are several orders of magnitude greater than the de-phasing length in the 1–2 keV spectral range, phase-matching techniques hold promise for increasing the current kiloelectronvolt photon flux of 10^2 – 10^3 photons per second (which is within 10% bandwidth at 1 keV) by several orders of magnitude. Radiation within this band is expected to emerge in a single pulse of less than 100 attoseconds when generated with a waveform-controlled, few-cycle pulse⁸. This would allow the investigation of a wide range of electronic dynamics in atoms, molecules and solids with near-atomic-scale resolution in time, and possibly simultaneously in space⁹.

J. Seres*, E. Seres*†, A. J. Verhoeff‡, G. Tempeas§, C. Strelli||, P. Wobrauschek||, V. Yakovlev¶, A. Scrinzi*, C. Spielmann†, F. Krausz*‡¶

*Institut für Photonik, Technische Universität Wien, 1040 Wien, Austria

e-mail: ferenc.krausz@mpq.mpg.de

†Physikalisches Institut EPI, Universität Würzburg, 97074 Würzburg, Germany

‡Max-Planck-Institut für Quantenoptik, 85748 Garching, Germany

§Femtolasers Produktions GmbH, 1100 Wien, Austria

||Atominstitut der Österreichischen Universitäten, Technische Universität Wien, 1020 Wien, Austria

¶Department für Physik, Ludwig-Maximilians-Universität München, 85748 Garching, Germany

- Macklin, J. J., Kmetec, J. D. & Gordon, C. L. *Phys. Rev. Lett.* **70**, 766–769 (1993).
- L'Huillier, A. & Balcou, P. *Phys. Rev. Lett.* **70**, 774–777 (1993).
- Seres, E., Seres, J., Krausz, F. & Spielmann, C. *Phys. Rev. Lett.* **92**, 163002 (2004).
- Gibson, E. A. et al. *Science* **302**, 95–98 (2003).
- Seres, J. et al. *Opt. Lett.* **28**, 1832–1834 (2003).
- Scrinzi, A., Geissler, M. & Brabec, T. *Phys. Rev. Lett.* **83**, 706–709 (1999).
- Lewenstein, M. et al. *Phys. Rev. A* **49**, 2117–2131 (1994).
- Kienberger, R. et al. *Nature* **427**, 817–820 (2004).
- Rischel, C. et al. *Nature* **390**, 490–492 (1997).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

brieF communications arising online

▶ www.nature.com/bca

Atmospheric science: Marine aerosols and iodine emissions

G. McFiggans (doi:10.1038/nature03372)

Reply: C. D. O'Dowd, J. L. Jimenez, R. Bahreini, R. C. Flagan, J. H. Seinfeld, K. Hämeri, L. Pirjola, M. Kulmala, S. G. Jennings, T. Hoffmann (doi:10.1038/nature03373)

Atmospheric science

Marine aerosols and iodine emissions

Arising from: C. O'Dowd *et al.* *Nature* **417**, 632–636 (2002)

O'Dowd *et al.*¹ describe the formation of marine aerosols from biogenic iodine and the growth of these aerosols into cloud-condensation nuclei (CCN). Based on chamber and modelling results, the authors suggest that biogenic organic iodine compounds emitted from macroalgae may be responsible for coastal particle bursts and that production of these compounds in the open ocean could increase CCN there too. It has since been shown that coastal particles are more likely to be produced from the photooxidation of molecular iodine². Moreover, I contend that open-ocean particle production and cloud enhancement do not result from emissions of organic iodine at atmospheric levels. For iodine particles to affect cloud properties over the remote ocean, an additional source of iodine is necessary as organic precursors cannot be responsible.

O'Dowd *et al.* propose that, in the presence of ozone, photolysis of atmospheric iodocarbons emitted from marine algae is a viable mechanism for the coastal production of new particles. This follows directly from observations relating aerosol formation and macroalgal iodocarbon emissions³ to the tidal cycle. However, particles formed by a variety of methods² have the same composition as those formed in the chamber study of O'Dowd *et al.*¹. The methods tested include exposing various *Laminaria* species to ambient and higher levels of ozone; mixing ozonized air with ozone-free air that has been passed over macroalgae; or exposing molecular iodine vapour (I_2) to ozone². The *Laminaria* macroalgae species are among those responsible for coastal iodine emissions.

Iodine-atom fluxes from an inorganic source, such as gaseous iodine⁴, have been shown to dominate² those from the organic compound diiodomethane (CH_2I_2 ; the most photolabile iodocarbon in the coastal environment and the one used by O'Dowd *et al.* in their chamber). Iodine-atom flux from atmospheric I_2 greatly exceeds that derived from the lowest CH_2I_2 concentrations used in the chamber experiments of O'Dowd *et al.*, which were more than an order of magnitude higher than atmospheric levels. Coastal iodine-containing particles are therefore much more likely to be formed from I_2 than from CH_2I_2 .

Particle bursts have been measured at Mace Head in Galway, Ireland, alongside iodocarbon fluxes. The phenomena are characterized by extremely high concentrations and production rates of 3-nm-diameter particles⁵. Some have deduced⁶ that the precursor vapour sources are biologically rich tidal regions, tens to hundreds of metres from the

measuring location (or tens of 'transit seconds'). On these timescales, average atmospheric CH_2I_2 concentrations, which are two orders of magnitude lower than those in the smog chamber¹, would not be able to form the observed 10^4 to 10^6 particles cm^{-3} . And yet for the laboratory mechanism to work in the coastal atmosphere, nucleation and growth characteristics must behave in the same way.

O'Dowd *et al.* used a model to demonstrate nanometre-scale sulphate-cluster stabilization by condensation of condensable iodine vapours (CIVs) and a consequent increase in CCN number. Their concentrations of CIVs are consistent with unrepresentatively low source estimates from a model⁷ that is aimed at explaining open-ocean iodine monoxide observations⁸. By assuming that CH_2I_2 is the coastal atmospheric iodine source, the model successfully replicates coastal observations⁷. However, coastal iodine-atom production from I_2 that is associated with an iodine monoxide mixing ratio of a few parts per trillion is about $5 \times 10^7 cm^{-3} s^{-1}$, and not $1 \times 10^4 cm^{-3} s^{-1}$ as calculated from CH_2I_2 .

Assuming a common chemical mechanism, the open-ocean iodine-atom production that is required to explain the observed remote marine iodine monoxide must be much higher than that proposed by O'Dowd *et al.* To achieve this rate, however, a further remote marine-atmospheric iodine source must be found (abiotic production⁹, for example), as the original flux estimate is consistent with the observed open-ocean organic iodine. At increased concentrations, iodine oxides may nucleate in the absence of pre-existing clusters^{1,2}, a mechanism that could explain the coastal phenomenon and possibly play a part in the open ocean. There is little evidence to support sulphate nucleation in the marine boundary layer following dimethylsulphide oxidation, hence the existence of pre-existing sulphate clusters should not be relied upon to explain marine-iodine enhancement of CCN. The identity of CIVs and their role in aerosol and CCN formation are unclear, as are their production rates as a function of an enhanced iodine-atom source. Taken together, all this calls for an improved understanding of the key species and processes involved in aerosol nucleation and growth in the marine boundary layer.

Gordon McFiggans

Atmospheric Science Group, School of Earth, Atmospheric and Environmental Sciences, University of Manchester, PO Box 88, Manchester M60 1QD, UK
e-mail: g.mcfiggans@manchester.ac.uk
doi:10.1038/nature03372

1. O'Dowd, C. *et al.* *Nature* **417**, 632–636 (2002).

2. McFiggans, G. *et al.* *Atmos. Chem. Phys.* **4**, 1–13 (2004).
3. Carpenter, L. J. *J. Geophys. Res.* **104**, 1679–1689 (1999).
4. Saiz-Lopez, A. & Plane, J. M. C. *Geophys. Res. Lett.* **31**, L04112 (2004).
5. O'Dowd, C. *et al.* *Geophys. Res. Lett.* **26**, 1707–1710 (1999).
6. O'Dowd, C. *et al.* *Atmos. Res.* **58**, 167–185 (2001).
7. McFiggans, G. *et al.* *J. Geophys. Res.* **105**, 14371–14385 (2000).
8. Allan, B. J., McFiggans, G., Plane, J. M. C. & Coe, H. J. *Geophys. Res.* **105**, 14361–14369 (2000).
9. Garland, J. A. & Curtis, H. J. *J. Geophys. Res.* **86**, 3183–3186 (1981).

O'Dowd *et al.* reply — McFiggans¹ raises some interesting, but partly speculative, issues about the possibility of additional condensable-iodine-vapour (CIV) precursors being involved in marine aerosol formation from biogenic iodine emissions, and about the relative roles of iodine oxide and sulphuric acid in the marine new-particle formation process.

We linked biogenic iodine compounds that were released by marine biota to particle formation through the formation of iodine oxides and focused on diiodomethane (CH_2I_2) as an iodine oxide precursor². As a result of new experiments, McFiggans now adds a further candidate to the list of aerosol precursors — molecular iodine (I_2). Like the iodocarbons, I_2 is rapidly photolysed during the daytime to yield highly reactive iodine atoms, which start the chain of gas-phase reactions that finally lead to the higher iodine oxides. This inclusion of another, previously unrecognized, iodine source reinforces our central hypothesis that biogenic iodine emissions have a potentially significant effect on global radiative forcing.

We agree that molecular iodine is a previously unidentified and important candidate that contributes to this process when macroalgae and tidal effects are considered: during the BIOFLUX campaign³, we measured I_2 and new particles at Mace Head in Galway, Ireland, at nearby biological hot-spots where nucleation intensity is even higher than at Mace Head, and in chambers containing different *Fucus* species. We found that there was a significant release of I_2 from macroalgae under natural conditions. During these experiments, the presence of I_2 was correlated with algal biomass, and the production of new particles was linearly correlated with I_2 concentrations; ambient concentrations of I_2 were of the order of 20–200 parts per trillion and so I_2 concentrations are higher than those previously reported for CH_2I_2 (about 1 part per trillion).

However, quantitative information on the source strength of I_2 is not available, nor is an understanding of the biological or chemical mechanism of its release. It is unknown which CIV source dominates over the open ocean, where different algal species are found and different sea–air transfer processes for iodine gases prevail. We highlighted² the fact that, based on the available information, there is a significant missing source or sources

of iodine over global oceans — hence the generic term CIV that we used is still valid.

McFiggans *et al.*⁴ link enhanced molecular iodine emissions to the significant concentration of inorganic iodine species in tidal waters as the water level subsides. This process is first linked to macroalgae as a source of precursors and second, and more particularly, to the concentrating effects of inorganic iodine species driven by tides that will not apply over the open ocean. Nevertheless, the suggestion that an additional source of iodine over the open ocean needs to be identified warrants further investigation.

The issue of nucleation of iodine oxides or thermodynamically stable sulphate clusters (TSCs) is still an open question. Detailed modelling studies⁵ of TSC nucleation in the marine boundary layer predict that TSC nucleation can easily occur under ambient conditions, but that these nuclei are rapidly scavenged by the existing aerosol before they can grow into aerosol particles or CCN. The addition of CIVs to the condensation process increases the survival probability of TSCs as they grow.

The suggestion that iodine concentrations could be greater than previously anticipated supports our predictions. But, given current understanding of the complex iodine-vapour system, it is not yet possible to determine whether CIV nucleation dominates over TSC nucleation. Whether or not CIVs nucleate and require sulphuric acid to grow nuclei into new aerosol particles, or vice versa, or whether a combination of both species participates in the nucleation and growth processes requires a deeper knowledge of gas-phase iodine processes and the role of iodine oxides in nucleation.

In conclusion, our original hypothesis remains valid but there is still an urgent need to understand more fully the production of iodine vapour, gas-phase interactions and particle production over the open ocean, as highlighted by McFiggans¹.

Colin D. O'Dowd*, **Jose L. Jimenez†**,
Roya Bahreini‡, **Richard C. Flagan‡**,
John H. Seinfeld‡, **Kaarle Hämeri§**,
Liisa Pirjolall, **Markku Kulmala¶**,
S. Gerard Jennings*, **Thorsten Hoffmann#**

*Department of Environmental Physics and

Environmental Change Institute, National University of Ireland, Galway, Ireland

†Department of Chemistry and Biochemistry, and CIRES, University of Colorado, Boulder, Colorado 80309-0216, USA

‡Departments of Environmental Science and Engineering and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, USA

§Finnish Institute for Occupational Health, Helsinki 00250, Finland

||Helsinki Polytechnic, Technology, Helsinki 00099, Finland

¶Department of Physical Sciences, University of Helsinki, Helsinki 00014, Finland

#Institute of Inorganic and Analytical Chemistry, Johannes Gutenberg-University of Mainz, Mainz 55128, Germany

doi:10.1038/nature03373

1. McFiggans, G. *Nature* doi:10.1038/nature03372 (2005).
2. O'Dowd, C. *et al. Nature* **417**, 632–636 (2002).
3. Yoon, Y. J. *et al. Proc. 16th Intl Conf. Nucleation and Atmospheric Aerosols*, Kyoto (eds Kasahara, M. & Kulmala, M.) 674–677 (Kyoto University Press, 2004).
4. McFiggans, G. *et al. Atmos. Chem. Phys.* **4**, 1–13 (2004).
5. Pirjola, L., O'Dowd, C. D., Brooks, I. M. & Kulmala, M. *J. Geophys. Res.* **105**, 26531–26546 (2000).

Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex

Kenichi Ohki, Sooyoung Chung*, Yeang H. Ch'ng*, Prakash Kara* & R. Clay Reid

Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, USA

* These middle authors contributed equally to this work

Neurons in the cerebral cortex are organized into anatomical columns, with ensembles of cells arranged from the surface to the white matter. Within a column, neurons often share functional properties, such as selectivity for stimulus orientation; columns with distinct properties, such as different preferred orientations, tile the cortical surface in orderly patterns. This functional architecture was discovered with the relatively sparse sampling of microelectrode recordings. Optical imaging of membrane voltage or metabolic activity elucidated the overall geometry of functional maps, but is averaged over many cells (resolution $>100\ \mu\text{m}$). Consequently, the purity of functional domains and the precision of the borders between them could not be resolved. Here, we labelled thousands of neurons of the visual cortex with a calcium-sensitive indicator *in vivo*. We then imaged the activity of neuronal populations at single-cell resolution with two-photon microscopy up to a depth of $400\ \mu\text{m}$. In rat primary visual cortex, neurons had robust orientation selectivity but there was no discernible local structure; neighbouring neurons often responded to different orientations. In area 18 of cat visual cortex, functional maps were organized at a fine scale. Neurons with opposite preferences for stimulus direction were segregated with extraordinary spatial precision in three dimensions, with columnar borders one to two cells wide. These results indicate that cortical maps can be built with single-cell precision.

One of the hallmarks of cortical functional architecture^{1,2} is that maps vary continuously across the cortical surface, but are punctuated by occasional jumps or discontinuities. Optical imaging of either metabolically related intrinsic signals or voltage-sensitive dyes^{3–5} has shown the two-dimensional structure of functional maps, including discontinuities in orientation⁶ and direction^{7,8} maps in visual cortex. Even in combination with electrophysiology⁹, however, it has been difficult to determine whether neurons of different functional types are mixed or kept separate near discontinuities. More generally, previous methods have been unable to examine whether there is any fine-scale organization—a functional micro-architecture—either near discontinuities or in cortical areas that lack functional architecture, such as rat visual cortex^{10,11}.

As an alternative to optical imaging of intrinsic signals or voltage-sensitive dyes, calcium-sensitive indicators¹² can be used to measure activity in single neurons with very high signal-to-noise ratios. When two-photon microscopy¹³ is used to image intracellularly loaded neurons in the cerebral cortex *in vivo*^{14,15} calcium signals in cell bodies and dendrites can be imaged with submicrometre resolution. Earlier experiments *in vitro*^{16,17} showed that bulk loading of neocortical cells with cell-permeant forms of calcium-sensitive indicators allows the activity of hundreds of neuronal cell bodies to be observed in parallel. We adapted a recent protocol for bulk loading *in vivo*¹⁸ to study the functional micro-architecture of visual cortex.

Functional architecture in rat visual cortex

We first examined the functional architecture of orientation selectivity in layer 2/3 of rat primary visual cortex. Several thousand layer 2/3 cells were loaded with a calcium indicator (Fig. 1a and Methods) injected into a region $200\text{--}400\ \mu\text{m}$ in diameter. Responses of labelled cells to visual stimuli were studied with gratings (light and dark bars) at four different orientations, each drifted in two opposite directions. These eight patterns were presented sequentially for 4–8 s each, interspersed with blank (uniform) stimuli of the

same duration. We imaged cells at a single focal plane continuously during multiple repetitions of the stimulus protocol. We repeated this protocol at multiple depths, $10\text{--}20\ \mu\text{m}$ apart, to study the three-dimensional functional architecture of layer 2/3 at single-cell resolution. At the end of these experiments, we often performed an anatomical reconstruction with $1\text{-}\mu\text{m}$ spacing in depth (Fig. 1a, top panel).

In rat primary visual cortex ($n=9$ animals), we found that 25–75% of stained cells at a given depth in layer 2/3 responded to at least one stimulus with a significant change in fluorescence ($P < 0.01$, analysis of variance (ANOVA) across blank and eight direction periods; see Methods). Of the visually responsive cells, 61% on average were selective for the orientation and/or direction of the stimulus ($P < 0.01$, ANOVA across eight directions). We found a smaller proportion of visually responsive cells compared with previous electrophysiological studies^{10,11,19}. This might be explained by our optimization of stimulus parameters for populations rather than individual neurons, and by our far more complete sampling of neurons (see Supplementary Discussion and Supplementary Fig. 1), some of which might not have been detected with electrophysiology.

As demonstrated elsewhere^{18,20}, somatic calcium signals reflect action potential activity, rather than subthreshold depolarizations. Our measurements were therefore sufficient to examine the relative selectivity of cells for different stimuli (see Supplementary Discussion). Because the drifting stimuli that we used typically evoke prolonged electrophysiological responses from layer 2/3 cortical neurons (see below), we could sample the calcium signal at a slow rate ($\sim 1\ \text{frames}^{-1}$).

The visually evoked calcium transients were analysed in several ways. First, we calculated single-condition maps of the change in fluorescence (ΔF) by averaging the images collected during each of the eight visual stimuli and subtracting the average baseline between stimuli (Fig. 1b, outer panels; see also Supplementary Fig. 2 for $\Delta F/F$ maps). The maps show cells (seen as light spots) that were active with specific orientations of visual stimuli. These data were com-

binned into a single, colour-coded orientation map (Fig. 1c), in which the hue of each pixel is determined by the best orientation and the colour saturation is proportional to the sharpness of orientation tuning²¹.

A cell-based analysis of our images allowed us to examine stimulus selectivity more quantitatively for individual neurons. First, cells in the anatomical image (Fig. 1a) were identified and outlined with a semi-automated algorithm. Next, all the pixels within a cell outline were summed to give a single time course ($\Delta F/F$) over the entire duration of the visual stimulation protocol (eight directions of motion). Single-cell averaged responses showed clear peaks in fluorescence (up to 30%) during visual stimulation (Fig. 1d, grey regions) and returned to baseline in the intervals between stimuli (Fig. 1d, white regions). These time courses are consistent with the sustained electrophysiological responses evoked by drifting-grating stimuli (see below) and the calcium kinetics of neurons (see Supplementary Discussion and Supplementary Fig. 1). The responses were often selective for stimulus orientation (Fig. 1d, cells 1–4) and sometimes for direction (cell 3). Selective responses could often be discerned in single trials (Fig. 1d). Neighbouring cells, for example cells 1–3 in Fig. 1e, were often highly tuned to

different orientations. Their responses were entirely uncorrelated (Fig. 1d), confirming that two-photon calcium imaging has single-cell resolution with little or no cross talk (see Supplementary Discussion). The orientation maps generated by the pixel-based and cell-based analyses are similar (Fig. 1c, e). The cell-based analysis allowed us to examine population statistics and is thus used in all subsequent figures.

Functional architecture in cat visual cortex

Previous single-unit electrophysiological studies have failed to discern any order in the orientation preference of nearby neurons in rat visual cortex^{10,11}. Even with our far more complete sampling at single-cell resolution, we found no discernible local structure (see Supplementary Discussion). By contrast, in cat visual cortex the overall functional architecture is exquisitely organized. For example, intrinsic-signal optical imaging has shown that area 18 of the cat visual cortex has regions with roughly constant orientation⁶ and direction preferences⁷ (for ferret, see ref. 8). These regions are punctuated by singularities in the orientation map (pinwheels⁶) and reversals in the direction map^{7,8}. The low resolution of optical imaging, however, could not show the degree of

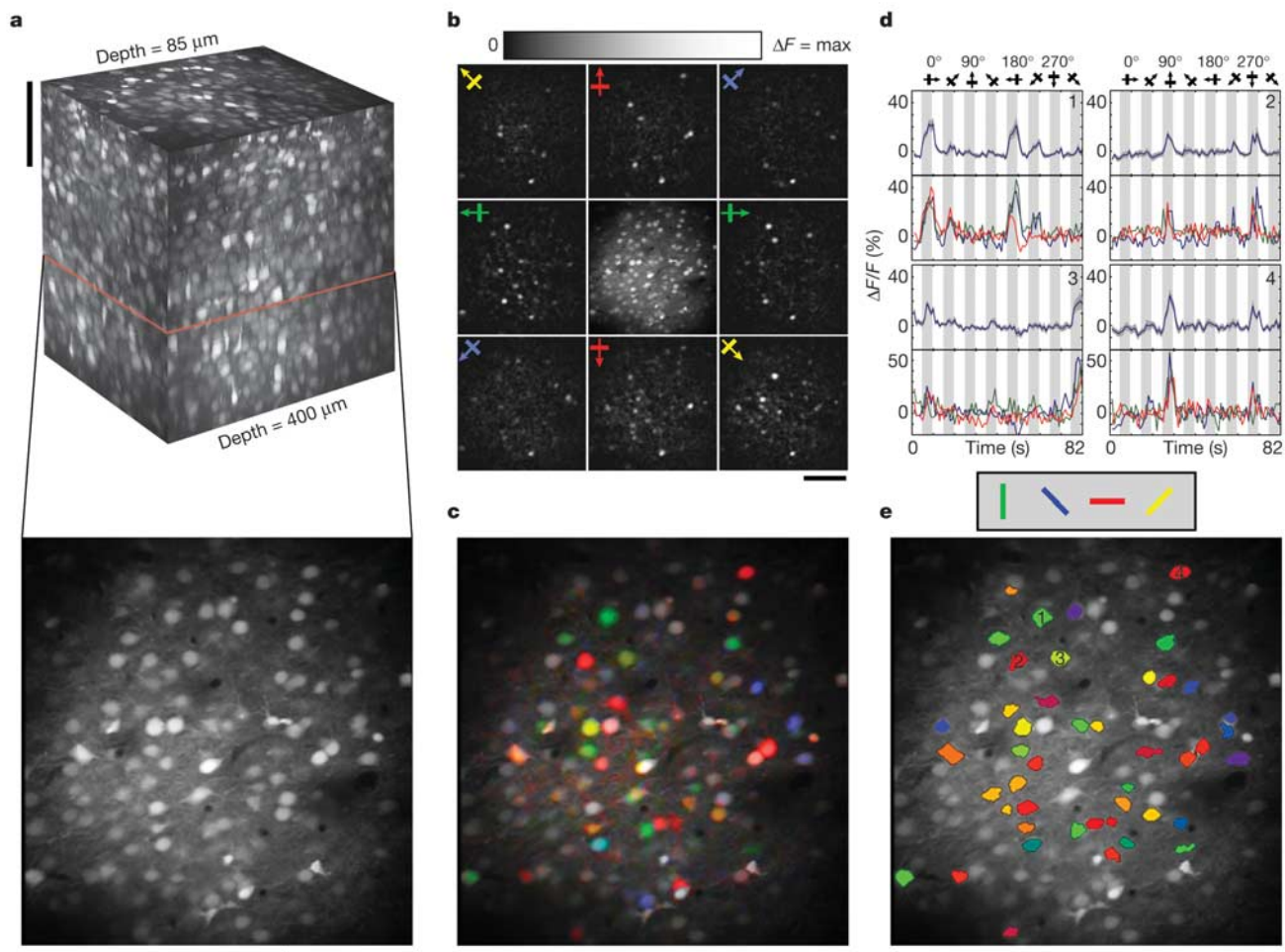


Figure 1 Functional maps of selective responses in rat visual cortex with single-cell resolution. **a**, *In vivo* images of cortical cells stained with a calcium indicator, OGB-1 AM. The top panel shows a volume of stained cells (depth of 85–400 μm including >3,000 cells) reconstructed from images obtained with 1 μm spacing in depth; the bottom panel is an anatomical image at 290 μm below the pia, averaged over all frames during the visual stimulation protocol. **b**, Single-condition maps (ΔF) for eight directions of visual stimuli (outer panels; the arrows indicate stimulus direction). Each map is the average of eight repeats. In this and subsequent figures, the scale bar (ΔF) only applies to the outer

panels. The central panel is re-drawn from **a**, **c**. Pixel-based orientation map, in which hue is determined by the best orientation (see text and Methods), overlaid with the anatomical image in **a**, **d**. Time courses of four orientation-selective cells (1–4) in **e**. The upper traces show the average response to eight repeats ($\pm$ s.e.m. in grey); lower traces show three individual repeats. Visual stimulation periods for eight directions are indicated by grey bars. **e**, Cell-based orientation map. Visually responsive cells ($P < 0.01$, ANOVA; 45/115) are coloured according to their preferred orientation. Scale bars, 100 μm.

uniformity of iso-direction or iso-orientation domains. It was therefore unknown whether all the neurons in a domain had similar responses, or some interspersed cells were tuned to very different orientations or directions. Furthermore, optical imaging could not be used to determine the spatial scale of either orientation singularities or the boundaries between directional domains, although electrophysiological studies have given some hints^{9,22,23}.

We therefore studied the functional micro-architecture of layer 2/3 in cat visual cortical area 18 ($n=10$ animals). Out of 6,734 cells identified, 63% were visually responsive ($P < 0.01$; ANOVA across blank and eight direction periods). Of the responsive cells, 97% were selective for orientation and/or direction ($P < 0.01$; ANOVA across eight directions). Of these, 86% were direction selective (direction index >0.33 ; see Methods for definition). One possible reason for the unresponsive cells (37%) is low signal-to-noise ratio in some dimly stained cells, particularly at the edges of the stained region. Another factor is that the stimulus parameters (bar size and drift velocity) were not optimized for all cells. However, it should be noted that in some cases, 100% of identified cells were responsive and tuned to orientation and/or direction.

In two out of ten experiments, we found regions where orientation and direction selectivity varied slowly over the entire field of view (a square region 240–300 μm on a side: smaller than the

400- μm spatial scale of direction or orientation domains^{6,7}). In one of these cases, single-condition maps (Fig. 2a, eight outer panels) showed that responses to 90° or 135° predominated. Of the 65 cells with statistically significant tuning for orientation or direction ($P < 0.01$), all had similar direction selectivity: no cells responded better to the opposite directions (270° and 315°; Fig. 2b). Furthermore, the transition of the preferred direction from 90° to 135° was smooth and ordered; each cell responded similarly to its nearest neighbours (Fig. 2b, c).

This uniformity and smoothness of direction maps leads to the question: what is the fine-scale structure near borders or discontinuities in the map? Because orientation pinwheels are sparsely scattered over cat visual cortical area 18 (refs 6, 24), we chose to examine the transition zones between regions of opposite direction preferences. The strict dichotomy between opposite directions offered an ideal opportunity to study the precision of a columnar border in the cerebral cortex. Optical imaging demonstrated overlap between regions that respond to opposite directions^{7,8}, but it could neither determine the spatial scale of the boundary between regions, nor indicate whether cells with opposite direction preferences are mixed together.

In eight out of ten experiments, our images included boundaries between regions of opposite direction selectivity. In one such case, single-condition direction maps (Fig. 3a) showed activation of cells (and neuropil to a lesser extent, see Supplementary Fig. 2) almost exclusively to stimuli of one orientation moving in either direction (45° and 225°). More importantly, the clusters of cells that were activated by opposite directions were highly segregated. To quantify the spatial scale of the direction border at single-cell resolution, we identified cells that were selective to orientation and colour coded them for their preferred direction (Fig. 3b). The time courses of single-cell responses (Fig. 3c) illustrate the reproducibility between individual trials and also the degree of direction tuning on both sides of the direction border.

Cells closer to the direction border were progressively less directionally selective. Nonetheless, almost all of these cells had a measurable bias so that a preferred direction could be assigned (Fig. 3b, grey–green and grey–red cells). The border between cells that were biased for opposite directions was so sharp that a straight line perfectly segregated them (vertical line in Fig. 3b). A second example from a different animal also shows a precise direction fracture (Fig. 3d).

The degree of segregation on either side of the direction border in Fig. 3b was quantified for three depths separated by 20 μm (Fig. 4; see also Supplementary Fig. 3) to examine the columnar border in three dimensions. At all three depths, the direction discontinuities (or fracture lines) were found to be precise at the scale of neighbouring cells: ~10–20 μm (Fig. 4a–c). As noted, neurons were progressively less directional (with direction index <0.33) in a transition zone 30–50 μm wide (Fig. 4d–f), but even the weakly biased cells had the same preferred direction as their neighbours on the same side of the border. The scales of the transition zone (30–50 μm) and particularly of the direction discontinuity (10–20 μm) are surprising given the 250–400- μm width of dendritic trees of neurons in cortical layer 2/3 (refs 25–27).

As a control, the direction tuning of calcium signals was compared with single-unit electrophysiology in two additional experiments. Calcium signals were imaged briefly at low laser power, followed by extensive electrophysiological recordings. In one of these experiments, calcium images showed both smoothly changing orientation domains and a direction discontinuity (Fig. 5a). Multiple penetrations with tungsten microelectrodes were made in different direction domains. Four representative cells responded in a sustained fashion to the drifting gratings for the entire period of stimulation (Fig. 5b, histograms). The direction selectivity of the electrophysiological responses matched the selectivity seen with calcium imaging in all four sites (Fig. 5b, polar plots),

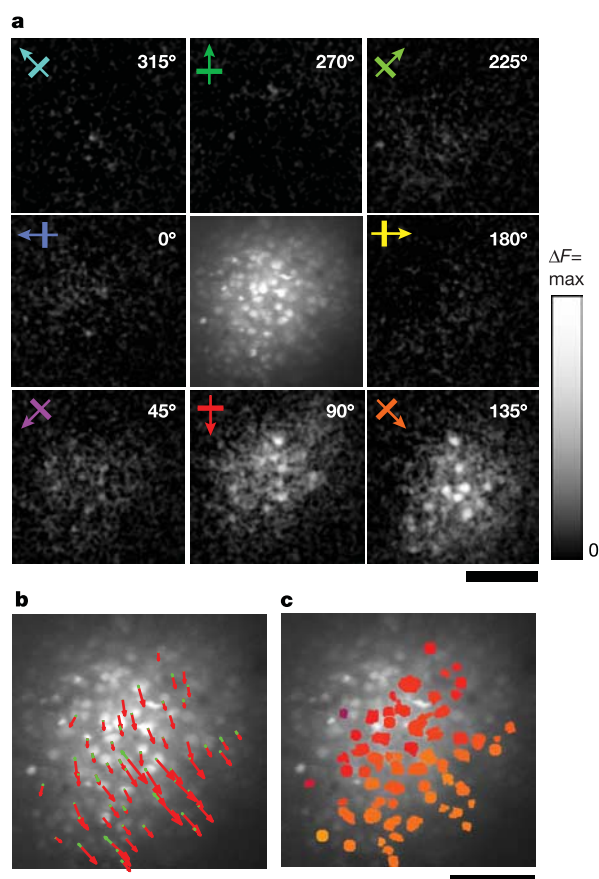


Figure 2 Smoothly changing direction map in cat visual cortex. **a**, Single-condition maps (ΔF) imaged 200 μm below the pia are shown in the outer panels. The central panel shows the anatomical image as in Fig. 1. Because of the wide brightness range, some stained cells are not discernible in this image. **b**, Vector map of preferred direction, calculated from cells with significant visual response and selectivity (per cent change $>5\%$ and $P < 0.01$, ANOVA). Red and green vectors indicate the magnitudes of the preferred and null responses, respectively. Preferred direction of identified cells changes smoothly (in a down and rightward direction) with no cells preferring upward motion. **c**, Colour-coded direction map, according to the colour keys in **a**. Scale bars, 100 μm .

including smoothly changing orientation domains (sites 1 and 3), weakly directional responses near the border (site 4), and opposite direction tuning across the border (site 2).

Three regimes of functional micro-organization

We have demonstrated three qualitatively different regimes of functional micro-organization in the visual cortex. The differences between these three regimes are illustrated in Fig. 6 for single experiments, but similar results were obtained for our entire sample (see Supplementary Discussion). First, in rat visual cortex the distance between cells had no relationship to their relative direction preferences (Fig. 6a; $r^2=0.005$, $P > 0.1$) or orientation preferences (not shown): highly tuned cells were distributed across the cortex with no apparent local structure (as in Fig. 1; but see also Supplementary Discussion). Second, in cat visual cortical area 18, direction domains are quite uniform; there are no interspersed cells tuned to the opposite directions (Fig. 2). Furthermore, the preferred direction varies smoothly at the level of neighbouring cells. Consequently, differences between preferred directions are positively correlated with the distance between cells (Fig. 6b; $r^2=0.16$, $P < 10^{-10}$). Third, we have found that when there are direction discontinuities in cat area 18, the borders are extra-

ordinarily sharp (Figs 3 and 4). In maps with discontinuities, cell pairs separated by any given distance have relative directional preferences that are bi-modally distributed (Fig. 6c). This bi-modal distribution follows from the sharp fracture between cells with opposite direction preferences (Fig. 4a–c).

The differences between the visual cortex in cats and rats raise the question of the role of functional architecture in cortical computations²⁸. As we have confirmed, cells in the rat can be orientation selective although there is no apparent functional architecture¹¹; however, cats²⁹ have narrower orientation and direction tuning than rats¹¹. One might therefore speculate that well-ordered cortical maps may provide an advantage in the sharpening of visual responses by connections between neighbouring neurons with similar stimulus selectivity.

Discussion

The ability to examine the receptive-field properties of almost every neuron in a local volume opens up a new realm of experimental investigations of the cerebral cortex. As illustrated in this study, it can show the degree of homogeneity and precision in cortical maps.

Maps that are locally disordered, such as the orientation map of rat visual cortex (Fig. 1), seem inconsistent with the theory that

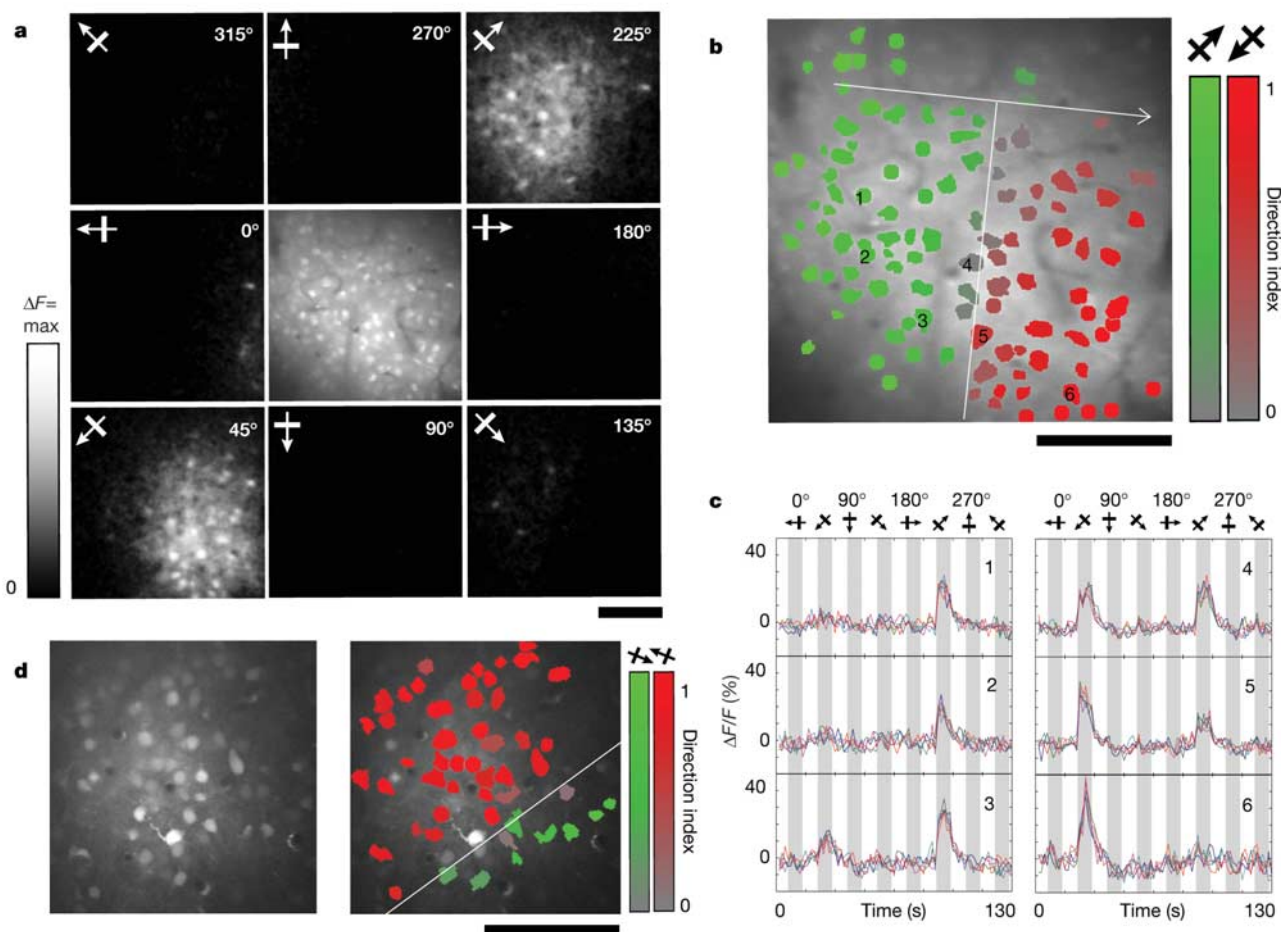


Figure 3 Direction discontinuity in cat visual cortex. **a**, Single-condition maps (ΔF) imaged 180 μm below the pia are shown in the outer panels. The central panel shows the anatomical image as in Fig. 1. Cells were activated almost exclusively by stimuli of one orientation moving in either direction (45° and 225°). To the non-preferred stimuli, such as 90°, the calcium responses were so small and the noise was so low that the single-condition maps are almost indistinguishable from zero. **b**, Cell-based direction map; 100% of cells had significant responses. Colour specifies preferred direction (green,

225°; red, 45°). The cells responding to both 45° and 225° are displayed as grey, according to their direction index (see colour scale on right). The vertical white line below the arrow indicates approximate position of the direction discontinuity. The discontinuity was curved in the upper right of the panel (see Supplementary Fig. 3), so we excluded the neurons above the arrow from quantitative analysis. **c**, Single-trial time courses of six cells, numbered (1–6) as in **b**. Five trials (out of ten) are superimposed. **d**, Direction discontinuity from another animal, imaged 240 μm below the pia. Scale bars, 100 μm .

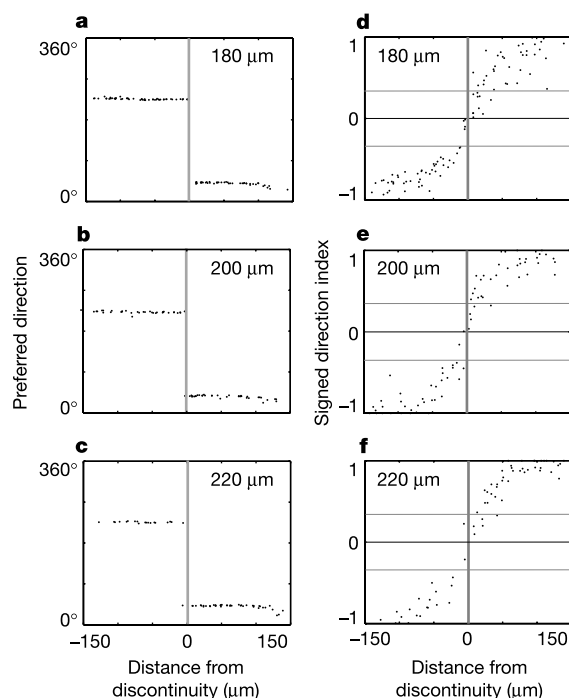


Figure 4 Sharpness of direction discontinuity at multiple depths. **a–c**, Preferred direction of each cell plotted against the distance from the direction discontinuity (vertical white line below arrow in Fig. 3b; see Supplementary Fig. 3). Single-cell precision was observed at three depths (**a**, 180 μm; **b**, 200 μm; **c**, 220 μm). The three discontinuity lines were shifted by 6 μm between sections, so they all fell in a near-vertical plane. **d–f**, The signed direction index (SDI, negative for 225°; positive for 45°) plotted against the distance from the direction discontinuity. Between two regions of high selectivity, the transition ($|SDI| < 0.33$; bounded by grey lines) was very rapid, in the range of 30–50 μm.

cortical connections are determined solely by the spatial extent of dendritic and axonal arborizations³⁰. Instead, connections onto cells with overlapping dendritic arborizations might be selective, as in the case of thalamic inputs to cat visual cortex³¹. For instance, groups of layer 2/3 neurons with similar orientation preferences might be selectively interconnected to form partially isolated subnetworks.

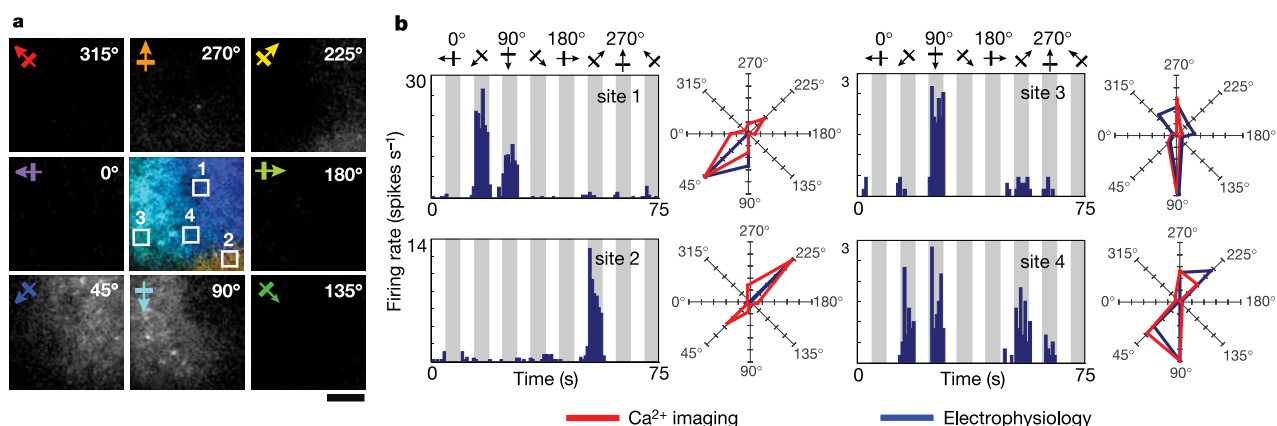


Figure 5 Correspondence of direction tuning obtained by calcium imaging and single-unit electrophysiology in cat visual cortex. **a**, Calcium imaging at 160 μm from the pia. Single-condition maps (ΔF) are shown in the outer panels. The central image is a hue-lightness-saturation map, in which preferred direction is colour coded according to the key shown as arrows in outer panels. **b**, Histograms are post-stimulus time histograms (PSTHs) of four single-unit responses, showing the average spike rates in response to the visual

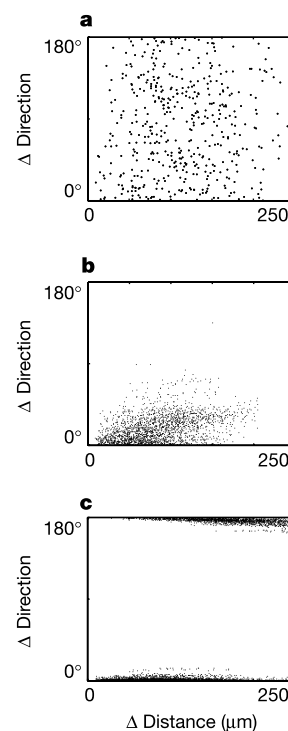


Figure 6 Three regimes of functional organization. For all pairs of cells, the difference in preferred direction is plotted against distance between them. **a**, Rat visual cortex (from Fig. 1e). **b**, A smoothly changing direction map in cat visual cortex (from Fig. 2c). **c**, A direction discontinuity in cat visual cortex (from Fig. 3b).

If maps are structured at a finer spatial scale than the dendritic trees of local neurons, as seen in the borders between direction columns in cat area 18, then what are the mechanisms driving this precision? There are several possibilities: (1) that selective connections between individual neurons with similar tuning dictate the receptive-field properties on either side of the border; (2) that smaller anatomical features, such as bundles of apical dendrites³², are functionally important; or (3) that precise inhibitory circuits might dominate neuronal response properties. Single-cell calcium

stimuli. Polar plots (to the right of PSTHs) show the normalized magnitude of responses to eight stimulus directions, obtained from calcium imaging (red) and single-unit activity (blue). Responses of single units were obtained from the PSTHs by summing firing during the stimulus period. Calcium signals were summed over a square region, 40 μm on a side (white boxes in **a**), which were selected to match the electrode positions, but not necessarily at the same depth.

imaging combined with other techniques might address these different possibilities. For example, stronger correlations between anatomy and physiology could be achieved by labelling cells to reveal their transmitters or other molecular markers, their projection patterns, or their detailed dendritic and axonal morphology.

From its inception, the idea of functional architecture has been considered both at a macroscopic scale, hundreds of micrometres to several millimetres, and at a microscopic scale, mini-columns $\sim 50\text{ }\mu\text{m}$ wide^{32,33}. Until now, the microscopic scale could only be studied with microelectrodes, which typically sample neurons sparsely and without precise localization. Single-cell calcium imaging *in vivo* promises to elucidate the relationship between anatomy and physiology in neuronal ensembles at an unprecedented level of completeness. Combined with a characterization of the sensory or motor function of single neurons, it could permit the study of micro-functional architecture throughout the cerebral cortex and reveal other examples of maps that have single-cell precision. □

Methods

Animal preparation

Long Evans rats (postnatal days 27–31) and cats (postnatal days 19–40) were prepared for *in vivo* two-photon imaging. Rats were anaesthetized with urethane (1.5 g kg^{-1} , intraperitoneally), supplemented by ketamine (20 mg kg^{-1}) during surgery. In cats, anaesthesia was induced with a mixture of ketamine (20 mg kg^{-1}) and acepromazine (0.2 mg kg^{-1} , intramuscularly), and continued with isoflurane (1–2% in surgery, 0.25–1% during imaging). Cats were paralysed with vecuronium bromide ($0.2\text{ mg kg}^{-1}\text{ h}^{-1}$, intravenously) and expired CO_2 regulated at 3.5–4.5% via mechanical ventilation. In rats and cats, a small craniotomy was made over the visual cortex, the dura reflected and the underlying cortex covered with agarose (3% in artificial cerebrospinal fluid, ACSF).

Dye loading and *in vivo* two-photon imaging

We modified the protocol of ref. 18 to load cortical neurons with a calcium-sensitive dye under continuous visual guidance through the two-photon microscope. A total of 0.8 mM Oregon Green 488 BAPTA-1 AM (OGB-1 AM) was dissolved in DMSO with 20% pluronic acid and mixed in ACSF containing 40 μM Alexa Fluor 594 (all from Molecular Probes). A patch pipette (tip diameter of 1 μm for rats, 4–6 μm for cats) was filled with this solution and inserted into the cortex to a depth of 200–370 μm from the surface. OGB-1 AM and Alexa Fluor 594 were pressure-ejected from the pipette (5–10 p.s.i. for 40–80 s). On the basis of the calibrations in ref. 18, we estimate our injection to be $<1\text{ pl}$, although we did not measure this directly. Because OGB-1 AM is weakly fluorescent before it becomes internalized in cells, the amount of OGB-1 AM ejection was inferred from Alexa Fluor 594, as visualized through the two-photon microscope. Full loading of the OGB-1 AM dye by cortical somata took 0.5–1 h in rats and 1.5–3 h in cats. After confirming loading, the pipette was withdrawn and the craniotomy sealed with a glass coverslip. Two-photon imaging of changes in calcium fluorescence in cortical cells was monitored with a custom-built microscope and a Tsunami (Spectra Physics) mode-locked Ti:sapphire laser (810 nm) or on a Leica TCS SP2 microscope coupled with a Mira 900 (Coherent Systems) mode-locked Ti:sapphire laser (810 nm). All of the cat data and some of the rat data were collected on the Leica system; the remaining rat data (including Fig. 1) were collected on the custom-built system. Excitation light was focused by $\times 40$ immersion objectives (0.8 numerical aperture). The average power delivered to the brain was $<45\text{ mW}$.

Visual stimulation, image data acquisition and electrophysiology

Drifting square-wave gratings (100% contrast, 1–2 Hz) were presented on a 14-inch LCD monitor at eight directions of motion in 45° steps. Spatial frequency was set at 0.025–0.05 cycles per degree in rats and 0.11–0.33 cycles per degree in cats. Each stimulus started with a blank period of uniform grey (4–8 s) followed by the same period of visual stimulation. The eight stimuli were presented sequentially and repeated five to ten times. Great care was taken to shield the microscope objective and the photomultipliers from stray light. Our observation that even neighbouring neurons responded to different orientations makes it unlikely that there were artefacts caused by stray light from the stimulus. Images were obtained by Leica software or the software developed by ref. 34. A square region of cortex 150–300 μm on a side was imaged at either 256×256 or 512×512 pixels at 0.81–1.63 s per frame. In all experiments, images were obtained at least from three depths separated by 10 or 20 μm . In some experiments, multiple images were obtained at 1- μm spacing for three-dimensional anatomical reconstruction. In two additional cats, we performed extracellular single-unit recordings with tungsten microelectrodes³⁵, after calcium imaging.

Data analysis

Images were analysed in Matlab (Mathworks) and ImageJ (National Institutes of Health). When the brain was drifting slowly, more than 2 μm (but not more than several micrometres) over >10 min of scanning time, images were realigned by maximizing the correlation between frames. Cells were automatically identified (1,992 cells from rats,

6,734 cells from cats) through a series of morphological filters that identified the contours of cell bodies based on intensity, size and shape. Cell outlines were visually inspected and the rare but clear errors were corrected manually. Time courses of individual cells were extracted by summing pixel values within cell contours. Occasionally, slow drift of the baseline signal over minutes was removed by a low-cut filter (cut-off, 2–4 min). Visually responsive cells were defined by ANOVA across blank and eight direction periods ($P < 0.01$; 737 cells from rats, 37%; 4,246 cells from cats, 63%). Of these, selective cells to orientation/direction were defined by ANOVA across eight direction periods ($P < 0.01$; 452 cells from rats, 61%; 4,113 cells from cats, 97%). In cell maps (Figs 2b, c and 3b, d), only selective cells are colour coded. For selective cells, the direction index was defined as $1 - (\text{response to null direction})/(\text{response to preferred direction})$. If the direction index was larger than 0.33 (ref. 36), the cell was considered to be direction selective (3,522 cells from cats, 86% of the total of orientation-selective cells). Preferred orientation was obtained by vector averaging³⁵. Preferred direction was defined as the angle that showed maximum response in the interpolated tuning curve³⁷. For pixel-based analysis, images were averaged over stimulus repetitions, and spatially smoothed (gaussian, $\sigma=1\text{ }\mu\text{m}$). Fluorescence change (ΔF) maps were obtained by subtracting images during the blank period from images during which one of eight directions was presented. In Fig. 1c, a hue-lightness-saturation orientation map²¹ was overlaid on the anatomical image.

Received 19 October; accepted 14 December 2004; doi:10.1038/nature03274.

Published online 19 January 2005.

- Mountcastle, V. B. Modality and topographic properties of single neurons of cat's somatic sensory cortex. *J. Neurophysiol.* **20**, 408–434 (1957).
- Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160**, 106–154 (1962).
- Grinvald, A., Anglister, L., Freeman, J. A., Hildesheim, R. & Manker, A. Real-time optical imaging of naturally evoked electrical activity in intact frog brain. *Nature* **308**, 848–850 (1984).
- Blasdel, G. G. & Salama, G. Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature* **321**, 579–585 (1986).
- Grinvald, A., Lieke, E., Frostig, R. D., Gilbert, C. D. & Wiesel, T. N. Functional architecture of cortex revealed by optical imaging of intrinsic signals. *Nature* **324**, 361–364 (1986).
- Bonhoeffer, T. & Grinvald, A. Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature* **353**, 429–431 (1991).
- Shmuel, A. & Grinvald, A. Functional organization for direction of motion and its relationship to orientation maps in cat area 18. *J. Neurosci.* **16**, 6945–6964 (1996).
- Weliky, M., Bosking, W. H. & Fitzpatrick, D. A systematic map of direction preference in primary visual cortex. *Nature* **379**, 725–728 (1996).
- Maldonado, P. E., Godecke, I., Gray, C. M. & Bonhoeffer, T. Orientation selectivity in pinwheel centers in cat striate cortex. *Science* **276**, 1551–1555 (1997).
- Parnavelas, J. G., Burne, R. A. & Lin, C. S. Receptive field properties of neurons in the visual cortex of the rat. *Neurosci. Lett.* **27**, 291–296 (1981).
- Girman, S. V., Sauve, Y. & Lund, R. D. Receptive field properties of single neurons in rat primary visual cortex. *J. Neurophysiol.* **82**, 301–311 (1999).
- Tsien, R. Y. Fluorescence measurement and photochemical manipulation of cytosolic free calcium. *Trends Neurosci.* **11**, 419–424 (1988).
- Denk, W., Strickler, J. H. & Webb, W. W. Two-photon laser scanning fluorescence microscopy. *Science* **248**, 73–76 (1990).
- Svoboda, K., Denk, W., Kleinfeld, D. & Tank, D. W. *In vivo* dendritic calcium dynamics in neocortical pyramidal neurons. *Nature* **385**, 161–165 (1997).
- Waters, J., Larkum, M., Sakmann, B. & Helmchen, F. Supralinear Ca^{2+} influx into dendritic tufts of layer 2/3 neocortical pyramidal neurons *in vitro* and *in vivo*. *J. Neurosci.* **23**, 8558–8567 (2003).
- Yuste, R. & Katz, L. C. Control of postsynaptic Ca^{2+} influx in developing neocortex by excitatory and inhibitory neurotransmitters. *Neuron* **6**, 333–344 (1991).
- Yuste, R., Peinado, A. & Katz, L. C. Neuronal domains in developing neocortex. *Science* **257**, 665–669 (1992).
- Stosiek, C., Garaschuk, O., Holthoff, K. & Konnerth, A. *In vivo* two-photon calcium imaging of neuronal networks. *Proc. Natl Acad. Sci. USA* **100**, 7319–7324 (2003).
- Wiesenfeld, Z. & Kornel, E. E. Receptive fields of single cells in the visual cortex of the hooded rat. *Brain Res.* **94**, 401–412 (1975).
- Mao, B. Q., Hamzei-Sichani, F., Aronov, D., Froemke, R. C. & Yuste, R. Dynamics of spontaneous activity in neocortical slices. *Neuron* **32**, 883–898 (2001).
- Ts'o, D. Y., Frostig, R. D., Lieke, E. E. & Grinvald, A. Functional organization of primate visual cortex revealed by high resolution optical imaging. *Science* **249**, 417–420 (1990).
- Payne, B. R., Berman, N. & Murphy, E. H. Organization of direction preferences in cat visual cortex. *Brain Res.* **211**, 445–450 (1981).
- Swindale, N. V., Matsubara, J. A. & Cynader, M. S. Surface organization of orientation and direction selectivity in cat area 18. *J. Neurosci.* **7**, 1414–1427 (1987).
- Ohki, K., Matsuda, Y., Ajima, A., Kim, D. S. & Tanaka, S. Arrangement of orientation pinwheel centers around area 17/18 transition zone in cat visual cortex. *Cereb. Cortex* **10**, 593–601 (2000).
- Gilbert, C. D. & Wiesel, T. N. Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex. *Nature* **280**, 120–125 (1979).
- Martin, K. A. & Whitteridge, D. The relationship of receptive field properties to the dendritic shape of neurones in the cat striate cortex. *J. Physiol. (Lond.)* **356**, 291–302 (1984).
- Hirsch, J. A. Synaptic physiology and receptive field structure in the early visual pathway of the cat. *Cereb. Cortex* **13**, 63–69 (2003).
- Adams, D. L. & Horton, J. C. Capricious expression of cortical columns in the primate brain. *Nature Neurosci.* **6**, 113–114 (2003).
- Rose, D. & Blakemore, C. An analysis of orientation selectivity in the cat's visual cortex. *Exp. Brain Res.* **20**, 1–17 (1974).
- Braitenberg, V. & Schuz, A. *Anatomy of the Cortex: Statistics and Geometry* (Springer, Berlin, 1991).
- Reid, R. C. & Alonso, J. M. Specificity of monosynaptic connections from thalamus to visual cortex. *Nature* **378**, 281–284 (1995).

32. Peters, A. & Yilmaz, E. Neuronal organization in area 17 of cat visual cortex. *Cereb. Cortex* **3**, 49–68 (1993).
33. Mountcastle, V. B. *Perceptual Neuroscience: the Cerebral Cortex* (Harvard University, Cambridge, 1998).
34. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2**, 13 (<http://www.biomedical-engineering-online.com>) (2003).
35. Kara, P. & Reid, R. C. Efficacy of retinal spikes in driving cortical responses. *J. Neurosci.* **23**, 8547–8557 (2003).
36. Berman, N. E., Wilkes, M. E. & Payne, B. R. Organization of orientation and direction selectivity in areas 17 and 18 of cat cerebral cortex. *J. Neurophysiol.* **58**, 676–699 (1987).
37. Kim, D. S., Matsuda, Y., Ohki, K., Ajima, A. & Tanaka, S. Geometrical and topological relationships between multiple functional maps in cat primary visual cortex. *Neuroreport* **10**, 2515–2522 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank E. Takahashi for involvement in the first set of experiments; B. Sabatini, W. Regehr, R. Yuste and F. Engert for discussions and technical advice; S. Yurgenson for technical support and programming; A. Kerlin and J. Leong for programming; A. Vagody for surgical assistance; and R. Yuste and J. Pezaris for comments on the manuscript. This work was supported by grants from the NEI and fellowships from the Uehara Foundation (K.O.), the Goldenson Fund (S.C.) and HHMI (Y.H.C.).

Authors' contributions K.O. started this work and played the major role in the project; S.C., Y.H.C. and P.K. contributed equally to its completion.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.C.R. (clay_reid@hms.harvard.edu).

Energy input from quasars regulates the growth and activity of black holes and their host galaxies

Tiziana Di Matteo^{1*}, Volker Springel¹ & Lars Hernquist²

¹Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Strasse 1, 85740 Garching bei München, Germany

²Astronomy Department, Harvard University, 60 Garden Street, Cambridge, Massachusetts 02138, USA

* Present address: Department of Physics, Carnegie-Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA

In the early Universe, while galaxies were still forming, black holes as massive as a billion solar masses powered quasars. Supermassive black holes are found at the centres of most galaxies today^{1–3}, where their masses are related to the velocity dispersions of stars in their host galaxies and hence to the mass of the central bulge of the galaxy^{4,5}. This suggests a link between the growth of the black holes and their host galaxies^{6–9}, which has indeed been assumed for a number of years. But the origin of the observed relation between black hole mass and stellar velocity dispersion, and its connection with the evolution of galaxies,

have remained unclear. Here we report simulations that simultaneously follow star formation and the growth of black holes during galaxy–galaxy collisions. We find that, in addition to generating a burst of star formation¹⁰, a merger leads to strong inflows that feed gas to the supermassive black hole and thereby power the quasar. The energy released by the quasar expels enough gas to quench both star formation and further black hole growth. This determines the lifetime of the quasar phase (approaching 100 million years) and explains the relationship between the black hole mass and the stellar velocity dispersion.

A large fraction of the black hole mass in galaxies today is thought to have been assembled during the peak of quasar activity in the early Universe^{11,12}, when large amounts of matter were available for accretion onto central black holes. Interactions and mergers between galaxies are known to trigger large-scale nuclear gas inflows^{13,14}, which are required for the growth of central black holes by accretion. Also, hierarchical models of galaxy formation imply that mergers of galaxies form elliptical or spheroidal components in galaxies, by destroying stellar disks and triggering nuclear starbursts.

This has led to suggestions that the $M_{\text{BH}}-\sigma$ relation (where M_{BH} is the black hole mass, and σ is the velocity dispersion of stars in the bulge of galaxies) could arise in galaxy mergers, provided that strong outflows are produced in response to major phases of accretion, capable of halting further black hole growth^{8,15–17}. Indeed,

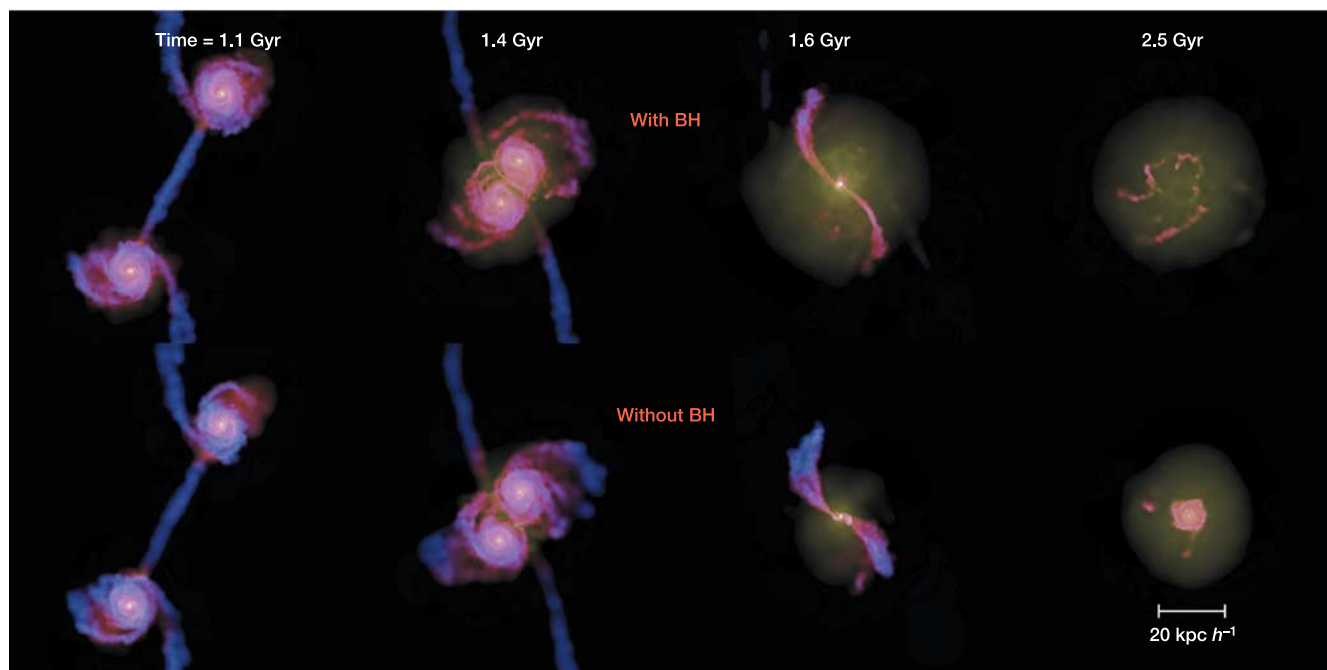


Figure 1 Snapshots of the simulated time evolution of mergers of two galaxies with and without black holes. The full time sequence for this simulation can be viewed in the Supplementary Video. Top and bottom panels show respectively the models with and without the inclusion of black holes (BH). In both cases, four snapshots at different times in the simulations are shown. The images visualize the projected gas distribution in the two galaxies, colour-coded by temperature (blue to red). The colliding galaxies have the same initial mass corresponding to a ‘virial velocity’ $V_{\text{vir}} = (M_{\text{tot}} \times 10 G H_0)^{1/3} = 160 \text{ km s}^{-1}$, and consist of an extended dark matter halo, a stellar bulge, and a disk made up of stars and 20% gas (here M_{tot} is the total halo mass, G is the gravitational constant and $H_0 = 100 \text{ h km s}^{-1} \text{ Mpc}^{-1}$ is the Hubble constant where we take the constant $h = 0.7$). Each individual galaxy in the simulations is represented with 30,000 particles for the dark matter, 20,000 for the stellar disk, 20,000 for the gaseous disk, and 10,000 for the bulge component. Two such galaxies were set up on a parabolic, prograde collision course, and then evolved forward in time numerically with GADGET-2 (ref. 29), a parallel tree-smooth-

particle-hydrodynamic simulation code. The first snapshot ($t = 1.1 \text{ Gyr}$) shows the systems after the first passage of the two galaxies. The second snapshot ($t = 1.4 \text{ Gyr}$) depicts the galaxies distorted by their mutual tidal interaction, just before they merge. The peak in the star formation and black hole accretion (see also Fig. 2) is reached at the time of the third snapshot ($t = 1.6 \text{ Gyr}$), when the galaxies finally coalesce. At this time, a strong wind driven by feedback energy from the accretion expels much of the gas from the inner regions in the simulation with black holes. Finally, the last snapshot shows the systems after the galaxies have merged ($t = 2.5 \text{ Gyr}$), leaving behind quasi-static spheroidal galaxies. In the simulation with black holes (top), the remnant is very gas poor and has little gas left that is dense enough to support ongoing star formation. This highlights the fact that the presence of supermassive black holes, which accrete matter from the surrounding gas and heat it with the associated feedback energy, dramatically alters the merger remnant. The scale bar shows the dimension in units of $\text{kpc } h^{-1}$ ($1 \text{ pc} = 3.26 \text{ light years}$).

observations point to the existence of strong outflows in bright quasars^{18–20}. In this picture, black hole accretion is expected to have a crucial effect on the evolution of the host galaxy. However, the coupling of star formation with black hole growth in the context of galaxy evolution is difficult to treat on the basis of analytical estimates alone.

We have, therefore, performed detailed numerical simulations of galaxy mergers that include radiative cooling, star formation, black hole growth, energetic feedback from supernovae and accretion onto black holes, as well as the gravitational dynamics of gas, stars and dark matter (see Supplementary Information and further details in ref. 21). As described in the Supplementary Information, we include a novel treatment of gas accretion onto supermassive black holes and its associated feedback in the centres of merging galaxies. We describe black holes using collisionless ‘sink’ particles that can grow in mass by accreting gas from their surroundings. The accretion rate is estimated by relating the small-scale, unresolved flow around the black hole to the large-scale, resolved gas properties using a spherical Bondi–Hoyle²² model. We further assume that a small fraction, f , of the radiated luminosity couples thermodynamically to the surrounding gas. This ‘feedback energy’ gives an effective feedback energy heating rate $\dot{E}_{\text{feed}} = fL = f\eta\dot{M}c^2$, where we fix the radiative efficiency, η , to 0.1, the typical value derived from Shakura–Sunyaev²³ accretion models onto a non-rotating black hole. (Here c is the velocity of light, $L = \eta\dot{M}c^2$ is the accretion luminosity, and $\dot{M}$ is the mass accretion rate). We further assume that $f = 0.05$, so that $\sim 0.5\%$ of the accreted rest mass energy is available to heat the gas.

To illustrate the effect of central, supermassive black holes on mergers of two disk galaxies, we compare the gas evolution between two simulations involving galaxies that are roughly the size of the Milky Way (Fig. 1). Star formation and supernova feedback are included in both simulations, but we add our black hole growth and feedback model in one (top four panels of Fig. 1; see also Supplementary Video) and neglect it in the other (bottom four panels). The first pair of images (at time $t = 1.1$ Gyr) shows the galaxies soon after their first encounter, when strong gravitational forces have spawned extended tidal tails. At this time, the black holes in the centres of each galaxy (top panel) have already grown significantly from their initial masses and are accreting at a moderate level. However, the overall star formation rate is essentially unaffected by the presence of the black holes, as indicated by Fig. 2, which plots the star formation rate (in both cases), the black hole accretion rate and black hole mass (for the top panels of Fig. 1) as a function of time.

The second snapshot ($t = 1.4$ Gyr) in Fig. 1 shows the galaxies when they begin to coalesce. Here, the tidal interaction has distorted the disks into a pair of bisymmetric spirals, and gas is shocked between the two galaxies. The tidal response drives gas into the central regions of each galaxy. Although only weak starbursts and accretion events are triggered at this time, it is already evident that black hole feedback alters the thermodynamic state of the gas, as indicated by the relatively lower density and higher temperature of the gas surrounding the galaxies in the simulation with black holes. In fact, a significant wind has started to flow out of the centres.

When the galaxies finally merge, as shown in the third pair of images in Fig. 1 ($t = 1.6$ Gyr), much of the gas is quickly converted into stars in intense bursts of star formation¹⁰. Owing to the enhanced gas density, the black holes, which also merge to form one object, experience a rapid phase of accretion close to the Eddington rate, resulting in significant mass growth. Also, the morphologies of the remnants in the two simulations begin to differ significantly. In the model without black holes, most of the gas is still inflowing in a comparatively cool phase. In contrast, the simulations with black holes exhibit a significant change in the thermodynamic state of the circumnuclear gas, which is heated by the feedback energy provided by the accretion and partly expelled in a powerful wind. During this strong accretion phase and for this

interval of time, the object would be a bright quasar with a specific lifetime.

Differences persist as the remnants settle into a relaxed state (fourth pair of images in Fig. 1, $t = 2.5$ Gyr). The remnant without black holes (bottom image) retains a large amount of dense cold gas, yielding prolonged star formation at a steady rate. However, in the simulation with supermassive black holes (top image), nearly all the gas is expelled from the centre, quenching star formation and black hole accretion itself. Consequently, the black hole mass saturates, quasar activity stops, and star formation is inhibited, so that the remnant resembles a ‘dead’ elliptical galaxy whose stellar population quickly reddens²⁴. In the particular example we show, the remnant of this major merger is an elliptical galaxy. In the hierarchical model of galaxy formation, a new disk can grow around this spheroid, turning it into the bulge component of a spiral galaxy. In very gas-rich mergers, a disk component may even survive directly²⁵ (an example of this is shown in the Supplementary Figure). We therefore expect black holes in bulges of spiral galaxies to be assembled in a manner similar to those in ellipticals. Our results should also apply for cases involving minor mergers.

The evolution of star formation rate, black hole accretion rate and black hole mass for mergers when the progenitor galaxy mass is varied (including the model shown in Fig. 1) are shown in Fig. 2. Models with different mass qualitatively reproduce the key features of the evolution shown in Fig. 1: the star formation and black hole accretion rates are both quenched in the remnant, and black hole growth saturates owing to feedback provided by accretion energy. However, the damping of star formation and black hole activity is

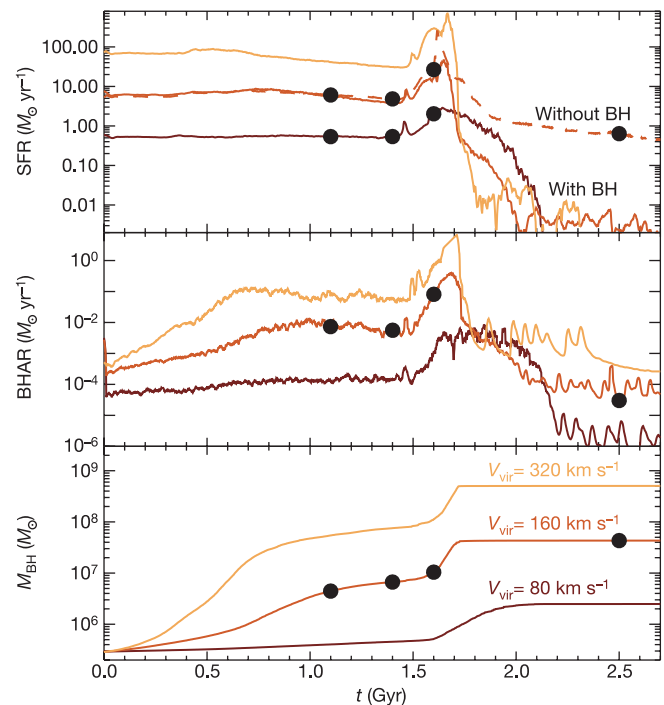


Figure 2 Black hole activity, star formation and black hole growth plotted as a function of time during a galaxy–galaxy merger. The star formation rate (SFR), black hole accretion rate (BHAR) and black hole mass are shown in the top, middle and bottom panels, respectively. The three lines in each panel (black, red and orange) correspond to models with galaxies of virial velocity $V_{\text{vir}} = 80, 160$ and 320 km s^{-1} , respectively. For comparison, we also show (top panel, dashed line) the evolution of the star formation rate for the model without a black hole that is shown in Fig. 1 (for the $V_{\text{vir}} = 160 \text{ km s}^{-1}$ galaxy). We note, in particular, that owing to AGN feedback, the peak amplitude of the starburst during the merger is lowered by a significant factor. The black filled circles in the individual panels identify the times of the corresponding snapshots shown in Fig. 1.

more abrupt in the more massive systems. Here, the total gas supply for accretion is larger, and the gravitational potential well is deeper, and so the black hole has to grow much more before its released energy is sufficient to expel the gas in a quasar driven wind, which then terminates further nuclear accretion and star formation. For the same reasons, the initial growth of the black holes, which is regulated by the properties of nearby gas, depends on the total mass. It is faster in more massive systems, which can therefore reach the exponential, Eddington-limited growth phase more easily. The lifetime of the active black hole phase, however, increases for smaller black hole masses, implying that low-luminosity quasars should be more numerous than bright ones. This is consistent with them residing in a greater number of smaller galaxies and with what has been found in recent surveys^{26,27}.

The dependence of black hole growth on galaxy mass yields a relation between the stellar spheroid of the remnant galaxy and its central black hole. Figure 3 shows the black hole mass versus the stellar velocity dispersion of the merger remnants from our simulations, compared with observations. We show simulations with six different galaxy masses, each of which has been run with three different initial gas mass fractions of the galaxies' disks. Remarkably,

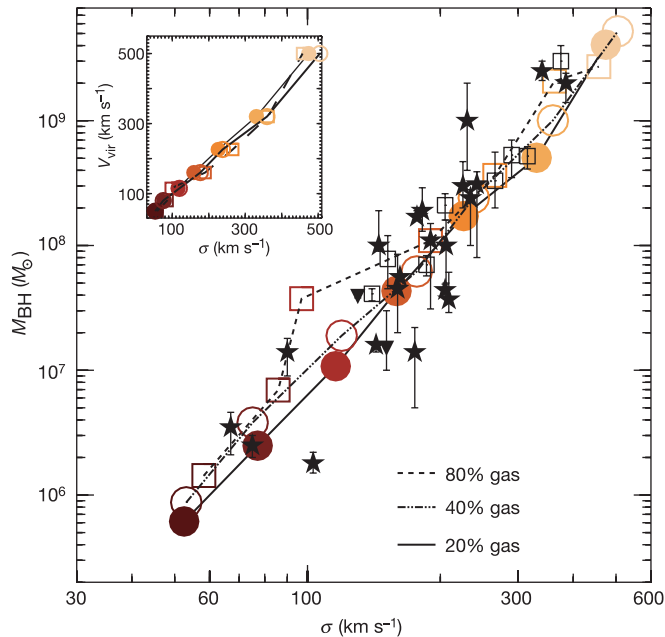


Figure 3 The relation between the final black hole mass, M_{BH} , and the velocity dispersion of stars, σ , of our galaxy merger simulations compared with observational measurements. Filled circles show the masses of the black holes and the bulge velocity dispersions measured for the final remnants of six merger simulations of galaxies with disk gas fraction of 20%, but different total mass, parameterized by virial velocities of $V_{\text{vir}} = 50, 80, 160, 320$ and 500 km s^{-1} (shown by the dark to light red, from low- to high-mass galaxies respectively). Open circles and open squares with the same colour give results for gas fractions of 40% and 80%, respectively. We have also checked that our results are insensitive to the orbits of the galaxy collisions. Mimicking the observational data, we calculate σ as the line-of-sight stellar velocity dispersion of stars in the bulge within the effective radius, R_e , of the galaxy. Black symbols show observational data for the masses of supermassive black holes and the velocity dispersions of their host bulges; measurements based on stellar kinematics are denoted by black filled stars, those on gas kinematics by black open squares, and those on maser kinematics by black filled triangles. Details for all the displayed measurements and associated r.m.s. error bars are given in refs 3 and 27. The observed BH sample has been fitted by a power law relation, yielding³⁰: $M_{\text{BH}} = (1.5 \pm 0.2) \times 10^8 M_{\odot} (\sigma/200)^{4.02 \pm 0.32}$ (here σ is in units of km s^{-1}). The inset shows the relation between the virial velocity, V_{vir} , and σ —shown in the x-axis of the main panel—measured for the merger remnants in the simulations. The same colour coding is used as in the main panel to indicate corresponding mass objects.

our simulations reproduce the observed $M_{\text{BH}}-\sigma$ correlation very well. Note that black holes in more gas-rich mergers reach somewhat larger masses than those growing in gas-poorer environments (which is expected from our prescription for the accretion rate), but this is partly compensated by an increase in the velocity dispersion of the corresponding bulges, maintaining a comparatively tight $M_{\text{BH}}-\sigma$ relation. However, this suggests that part of the intrinsic scatter in the observed relation can be ascribed to different gas fractions of the galaxies during black hole growth. Our lowest mass galaxy models probe a region of the $M_{\text{BH}}-\sigma$ relation where few measurements are available, and predict that this correlation should hold towards small black hole masses and small velocity dispersions, in tentative agreement with recent observations²⁸.

Black hole growth is self-regulated in our models. As galaxies merge to form spheroids, the dynamical response of the gas to the energy supplied by accretion halts further growth once the black holes have reached a critical size for the gravitational potential of the bulge. At this saturation point, the active galactic nuclei (AGN) generate outflows that drive away gas and inhibit further star formation. Our simulations are, to our knowledge, the first self-consistent models to demonstrate that self-regulation can quantitatively account for the principle observational facts known for the local population of supermassive black holes, most notably the $M_{\text{BH}}-\sigma$ relation. Moreover, self-regulation in our hydrodynamical simulations predicts a specific duration of the luminous episode of a black hole in a given galaxy, thereby explaining the origin of quasar lifetimes. We note that the final black hole masses we obtain are roughly proportional to the inverse of the value assumed for the feedback efficiency, f . We note that our choice of $f = 0.05$ is consistent with the value required in semi-analytic models⁸ to explain the evolution of the number density of quasars.

The black hole accretion activity also has a profound effect on the host galaxy. The remnant spheroid is gas-poor and has low residual star formation, so it evolves to a red stellar colour on a short timescale. The simulations shown here make it possible to draw firm conclusions about this and other links between black hole growth, quasar activity and properties of the galaxy population. Our approach can also be implemented in cosmological simulations of hierarchical structure formation in representative pieces of the Universe. Such simulations will allow us to study directly why quasars were much more numerous in the early Universe than they are today, and how black holes and galaxies have influenced each other throughout cosmic history. □

Received 23 September; accepted 21 December 2004; doi:10.1038/nature03335.

- Kormendy, J. & Richstone, D. Inward bound—The search for supermassive black holes in galactic nuclei. *Annu. Rev. Astron. Astrophys.* **33**, 581–624 (1995).
- Magorrian, J. et al. The demography of massive dark objects in galaxy centers. *Astron. J.* **115**, 2285–2305 (1998).
- Ferrarese, L. & Ford, H. C. Supermassive black holes in galactic nuclei: Past, present and future research. *Space Sci. Rev.* (in the press).
- Ferrarese, L. & Merritt, D. A Fundamental relation between supermassive black holes and their host galaxies. *Astrophys. J.* **539**, L1–L4 (2000).
- Gebhardt, K. et al. A relationship between nuclear black hole mass and galaxy velocity dispersion. *Astrophys. J.* **539**, L13–L16 (2000).
- Kauffmann, G. & Haehnelt, M. A unified model for the evolution of galaxies and quasars. *Mon. Not. R. Astron. Soc.* **311**, 576–588 (2000).
- Volonteri, M., Haardt, F. & Madau, P. The assembly and merging history of supermassive black holes in hierarchical models of galaxy formation. *Astrophys. J.* **582**, 559–573 (2003).
- Wyithe, J. S. B. & Loeb, A. Self-regulated growth of supermassive black holes in galaxies as the origin of the optical and X-ray luminosity functions of quasars. *Astrophys. J.* **595**, 614–623 (2003).
- Granato, G. L., De Zotti, G., Silva, L., Bressan, A. & Danese, L. A Physical model for the coevolution of QSOs and their spheroidal hosts. *Astrophys. J.* **600**, 580–594 (2004).
- Mihos, J. C. & Hernquist, L. Gasdynamics and starbursts in major mergers. *Astrophys. J.* **464**, 641–663 (1996).
- Soltan, A. Masses of quasars. *Mon. Not. R. Astron. Soc.* **200**, 115–122 (1982).
- Yu, Q. & Tremaine, S. Observational constraints on growth of massive black holes. *Mon. Not. R. Astron. Soc.* **335**, 965–976 (2002).
- Hernquist, L. Tidal triggering of starbursts and nuclear activity in galaxies. *Nature* **340**, 687–691 (1989).
- Barnes, J. & Hernquist, L. Dynamics of interacting galaxies. *Annu. Rev. Astron. Astrophys.* **30**, 705–742 (1992).
- Silk, J. & Rees, M. Quasars and galaxy formation. *Astron. Astrophys.* **334**, L1–L4 (1998).

16. Fabian, A. C. The obscured growth of massive black holes. *Mon. Not. R. Astron. Soc.* **308**, L39–L43 (1999).
17. King, A. Black holes, galaxy formation, and the $M_{\text{BH}}-\sigma$ relation. *Astrophys. J.* **596**, L27–L29 (2003).
18. Chartas, G., Brandt, W. N. & Gallagher, S. C. XMM-Newton reveals the quasar outflow in PG 1115+080. *Astrophys. J.* **595**, 85–93 (2003).
19. Crenshaw, D. M., Kraemer, S. B. & George, I. M. Mass loss from the nuclei of active galaxies. *Annu. Rev. Astron. Astrophys.* **41**, 117–167 (2003).
20. Pounds, K. A. *et al.* A high-velocity ionized outflow and XUV photosphere in the narrow emission line quasar PG1211+143. *Mon. Not. R. Astron. Soc.* **345**, 705–713 (2003).
21. Springel, V., Di Matteo, T. & Hernquist, L. Modeling feedback from stars and black holes in galaxy mergers. *Mon. Not. R. Astron. Soc.* (submitted); Preprint astro-ph/0411108 at (<http://xxx.lanl.gov/>) (2004).
22. Bondi, H. On spherically symmetrical accretion. *Mon. Not. R. Astron. Soc.* **112**, 195–204 (1952).
23. Shakura, N. I. & Sunyaev, R. A. Black holes in binary systems. Observational appearance. *Astron. Astrophys.* **24**, 337–355 (1973).
24. Springel, V., Di Matteo, T. & Hernquist, L. Black holes in galaxy mergers: The formation of red elliptical galaxies. *Astrophys. J.* (submitted).
25. Springel, V. & Hernquist, L. Formation of a spiral galaxy in a major merger. *Astrophys. J.* (submitted).
26. Hasinger, G., Miyaji, T. & Schmidt, M. Luminosity dependent evolution of soft X-ray selected AGN. *Astron. Astrophys.* (submitted).
27. Barger, A. J. *et al.* The cosmic evolution of hard X-ray selected active galactic nuclei. *Astrophys. J.* (in the press).
28. Greene, J. E. & Ho, L. C. Active galactic nuclei with candidate intermediate-mass black holes. *Astrophys. J.* (in the press).
29. Springel, V. & Hernquist, L. Cosmological smoothed particle hydrodynamics simulations: a hybrid multiphase model for star formation. *Mon. Not. R. Astron. Soc.* **339**, 289–311 (2003).
30. Tremaine, S. *et al.* The slope of the black hole mass versus velocity dispersion correlation. *Astrophys. J.* **574**, 740–753 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The computations reported here were performed at the Center for Parallel Astrophysical Computing at the Harvard-Smithsonian Center for Astrophysics and at the Rechenzentrum der Max-Planck-Gesellschaft in Garching.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.D.M. (tiziana@phys.cmu.edu).

Electronically soft phases in manganites

G. C. Milward, M. J. Calderón & P. B. Littlewood

Cavendish Laboratory, Cambridge University, Madingley Road, Cambridge CB3 0HE, UK

The phenomenon of colossal magnetoresistance in manganites¹ is generally agreed to be a result of competition between crystal phases with different electronic, magnetic and structural order; a competition which can be strong enough to cause phase separation between metallic ferromagnetic and insulating charge-modulated states^{2–5}. Nevertheless, closer inspection of phase diagrams in many manganites reveals complex phases where the two order parameters of magnetism and charge modulation unexpectedly coexist^{6,7}. Here we show that such experiments can be naturally explained within a phenomenological Ginzburg–Landau theory. In contrast to models where phase separation originates from disorder⁸ or as a strain-induced kinetic phenomenon⁹, we argue that magnetic and charge modulation coexist in new thermodynamic phases. This leads to a rich diagram of equilibrium phases, qualitatively similar to those seen experimentally. The success of this model argues for a fundamental reinterpretation of the nature of charge modulation in these materials, from a localized to a more extended ‘charge-density wave’ picture. The same symmetry considerations that favour textured coexistence of charge and magnetic order may apply to many electronic systems with competing phases. The resulting

‘electronically soft’ phases of matter with incommensurate, inhomogeneous and mixed order may be general phenomena in correlated systems.

The manganese perovskites ($\text{RE}_{1-x}\text{AE}_x\text{MnO}_3$, where RE is rare earth and AE is alkaline earth) provide a laboratory in which to study the interplay of a variety of magnetic, electronic and structural phases of matter in a strongly correlated electronic system. As in many strongly correlated electronic systems, the basic paradigm for manganite physics is the competition between the delocalizing effects of the electron kinetic energy and the localizing effects of the Coulomb repulsion, aided by coupling to lattice degrees of freedom. When the kinetic energy is dominant, one finds a metallic ground state with ferromagnetic alignment of the core moments. When the localizing effects are dominant, we see instead charge and orbitally ordered ground states with substantial local lattice distortions from the near-cubic symmetry of the metal, along with insulating behaviour and antiferromagnetism. We can tune between these two phases using any of several external parameters, especially chemical substitution, but also lattice strain and magnetic field. The competition between metal and insulator is evident in the phenomenon of bulk colossal magnetoresistance, in which a magnetic field tunes the conductivity of the material, and is even clearer in the strong tendency toward phase separation and inhomogeneity and regimes of percolative transport.

The origin of charge and orbital ordered phases is still the subject of debate. Charge modulation has been traditionally seen as the ordering of Mn^{4+} and Mn^{3+} ions¹⁰. More recently, the charge disproportionation of the Mn ions has been argued to be much smaller than one^{11–13} but the idea of two kinds of cation-forming stripes still prevails and was used to interpret the experiments. In such a scenario, we would expect a density x of one kind of cation and $1 - x$ of the other. The charge modulation would be given by the averaged wave vector $\mathbf{q} \approx (1 - x)\mathbf{a}^*$, where $\mathbf{a}^*$ is the reciprocal lattice vector, aside from possible commensuration effects near special dopings ($x = 1/2, 2/3, 3/4$).

However, this picture is not compatible with the experimental findings on commensurate and incommensurate modulation that are summarized in Fig. 1a. At half-doping, commensurate modulation is expected and found at low temperatures, together with antiferromagnetism of the CE type¹⁰. Above the Néel temperature, though, the modulation is incommensurate^{7,14}. When found to be charge-modulated, underdoped samples ($x < 0.5$) do not show the relation $\mathbf{q} \approx (1 - x)\mathbf{a}^*$; instead the modulation wave vector is always commensurate with $\mathbf{q} = 0.5 \mathbf{a}^*$ (refs 15–17) and independent of temperature. Finally, the overdoped samples ($x > 0.5$) show the expected incommensurate wave vector¹⁸ below the Néel temperature, decreasing above it¹⁹. However, no sign of discommensurations, only a uniform incommensurate modulation, has been found in recent experiments on $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ (ref. 20).

Another experimental conundrum is the coexistence of charge modulation and ferromagnetism despite their natural antipathy. For instance, at half-doping, the incommensurate modulation above the Néel temperature is accompanied by ferromagnetism⁷. A different electronic phase showing ferromagnetism and charge modulation has also been found at low temperature in $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ (ref. 6). Slightly overdoped samples can also be ferromagnetic above the Néel temperature²¹. Another example of coexistence is given by the underdoped ($0.3 < x < 0.5$) $\text{Pr}_{1-x}\text{Ca}_x\text{MnO}_3$. The ground state is commensurate and charge-modulated^{15,16} but the antiferromagnetism is canted^{17,22}, showing a ferromagnetic component coexisting with the commensurate charge modulation.

We might attempt to explain these phenomena in terms of microscopic theory, incorporating the many different couplings and microscopic degrees of freedom; this is a daunting task, however, that might not be illuminating, owing to its complexity. Here we propose a simple and more transparent phenomenological approach to this problem, using the Ginzburg–Landau theory^{23–25}.

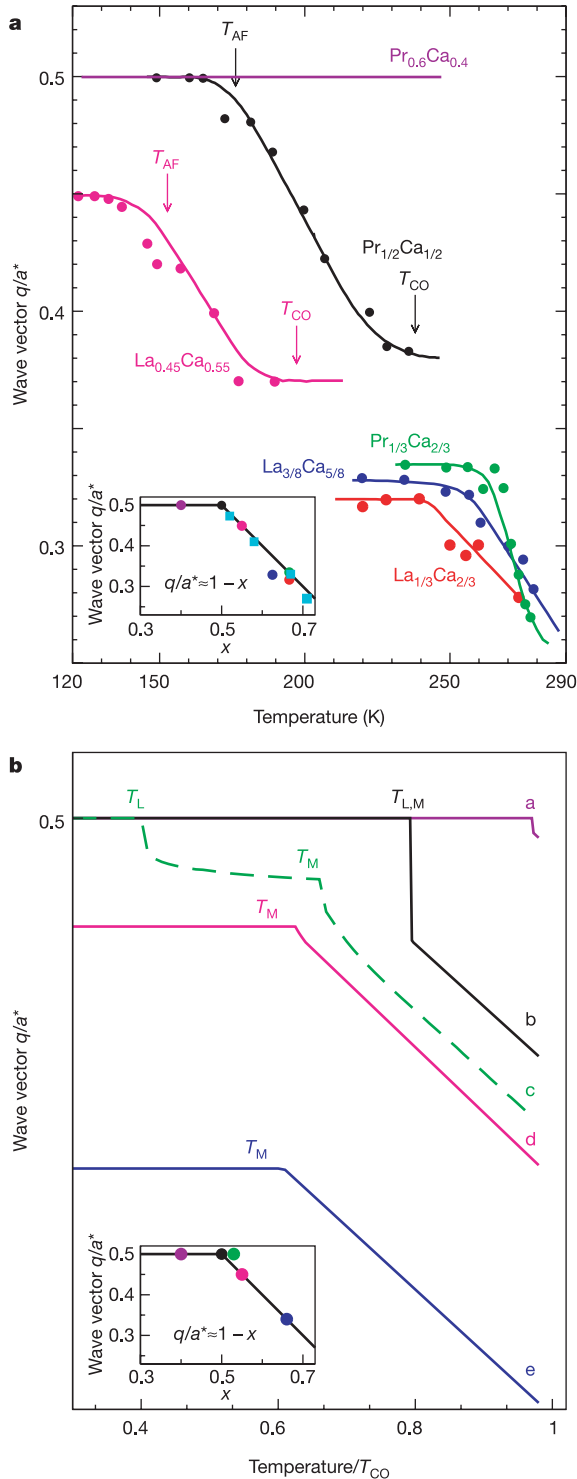


Figure 1 Wave vector of the modulation q/a^* versus temperature. **a**, Compilation of experimental data for $\text{Pr}_{1-x}\text{Ca}_x\text{MnO}_3$ and $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ at different dopings ($x \geq 0.5$ data taken from Figs 1 and 2 in ref. 19, and $x = 0.4$ taken from ref. 15). The same kind of behaviour has been reported for $\text{La}_{1/2}\text{Ca}_{1/2}\text{MnO}_3$ (ref. 7; not shown here). At low temperatures $q/a^* \approx (1 - x)$, as shown in the inset, and decreases above the Néel temperature T_{AF} . The inset shows q/a^* versus x as in refs 16 and 18. The circles correspond to the low-temperature values of the curves in the main panel, while the squares are data for $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$, taken from ref. 20. **b**, Generic curves as given by the present model. The lettered curves correspond to cuts in the phase diagram as specified in Fig. 2. T_L is the incommensurate to commensurate order transition and T_M is the temperature below which the magnetization is zero. The transition at T_M is first order, although very weakly so for large x . There is no lock-in transition T_L in the overdoped region.

We will show how the close interplay of ferromagnetism and charge modulation conspires to reproduce the experimental findings just discussed.

Ginzburg–Landau theory allows the study of phase transitions in a phenomenological way and it consists in expressing the free energy as a power expansion of the order parameters and their gradients. The order parameters we consider here are the magnetization $M(r)$ and the charge-orbital modulation $\psi(r) = \rho(r)e^{i(Q_c r + \phi(r))}$. r is the spatial coordinate, ρ is the amplitude of the modulation, $Q_c = a^*/n$ is a wave vector commensurate with the lattice and ϕ is the phase that would incorporate structures with incommensurate periodicities. To simplify the discussion we study a one-dimensional scalar modulation, because charge modulation within a domain occurs only in one direction. $n = 4$ gives the correct periodicity for the lattice distortions measured in $x = 0.5$ because, although the charge modulation has period 2, the orbital order follows a zig-zag pattern with period 4. Notice that if $\nabla\phi = 0$, ψ is a wave of amplitude ρ and wave vector commensurate with the lattice. If $\nabla\phi \neq 0$, the wave vector is $Q_c + \langle \nabla\phi \rangle$ and therefore, in general, cannot be expressed as a simple rational number of a^* .

The free-energy density can be separated into three contributions: magnetization, charge modulation and coupling terms. The first two are:

$$\mathcal{F}_M = \frac{1}{2}a_M(T - T_c)M^2 + \frac{1}{4}b_MM^4 + \frac{1}{2}\xi_M^2(\nabla M)^2 \quad (1)$$

$$\mathcal{F}_\psi = \frac{1}{2}a_\rho(T - T_{CO})\rho^2 + \frac{1}{4}b_\rho\rho^4 + \frac{1}{2}\xi_\rho^2(\nabla\rho)^2 + \frac{1}{2}\xi_\rho^2\rho^2(\nabla\phi - q_0)^2 + \frac{1}{n}\eta\rho^n\cos(n\phi) \quad (2)$$

The magnetic energy $\mathcal{F}_M$, taken alone, will describe a phase transition to homogeneous magnetism below the Curie temperature T_c . $\mathcal{F}_\psi$ is the free energy extensively used to study commensurate–incommensurate phase transitions of charge density waves, spin density waves or modulated lattice distortions²³. $q_0 = 1/2 - x$ is the predicted deviation (by chemical composition) from commensurability around $x = 0.5$. The term $\frac{1}{2}\xi_\rho^2\rho^2(\nabla\phi - q_0)^2$ favours a uniform incommensurate modulation with $\nabla\phi = q_0$. On the other

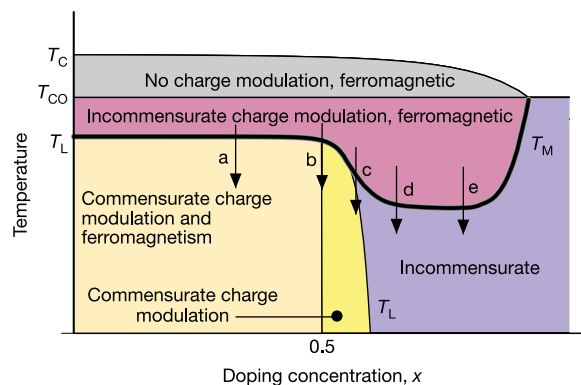


Figure 2 Schematic phase diagram that results from the minimization of the free energy. The scale of the axes depends on the particular parameters used. The commensurate order phase just above $x = 0.5$ has not been observed but we predict that it can be relevant for dopings very close to $x = 0.5$ for highly insulating manganites. The complex phases arise provided $T_c > T_{CO}$, a condition that can be relaxed if the model is extended to account for discontinuous phase transitions at T_{CO} . The values of T_{CO} and T_c are direct parameters in the model, whereas the lock-in temperature T_L is a consequence of the competition between the Umklapp and incommensurate modulation terms. T_M is the temperature below which the magnetization is zero. The thick solid line indicates the first-order transition and the thin lines indicate second-order transitions. The labels a to e correspond to the placement of curves in Fig. 1b.

hand $\frac{1}{n}\eta\rho^n \cos(n\phi)$ is an Umklapp term that favours commensurability with $\phi = 2\pi j/n$, where j is an integer (for $\eta < 0$). Taken alone, this describes two phases. Upon cooling below T_{CO} , the amplitude ρ of the charge density wave (CDW) rises from zero, but provided $n > 2$, the Umklapp term is small and the modulation is incommensurate. As temperature is lowered, ρ grows, the Umklapp term may become dominant and a lock-in transition occurs if η is comparable to ξ_ρ^2 .

We now discuss coupling between the two order parameters. The lowest-order coupling term which arises is $d_1\rho^2M^2$ so that there is a free-energy penalty for homogeneous coexistence of magnetism and charge modulation. Were this the only coupling term, the free energy would be generally stabilized either by a homogeneous magnetization or by charge modulation, depending on which transition temperature is the larger. Next, we can of course introduce uniform coupling terms of higher powers of M and ρ , but they make no qualitative changes unless they have a negative sign. More interesting is that there is a leading order coupling term in the gradient, not included in previous models^{24,25}, of the form $d_2\rho^2M^2(\nabla\phi - q_0)$. That there is a linear term in the gradient is expected because there is no symmetry about $x = 1/2$; different signs of the gradient correspond physically to compression or extension of the CDW period, that is, to extra '3+' or '4+' sites. We can also justify this term microscopically. If we consider the effect of charge modulation on the Fermi surface, then it is clear that by choosing a wave vector that does not match the chemical doping, we will be left with small pockets of carriers at the Fermi surface; these metallic electrons (or holes) are then available to mediate double exchange and thereby promote ferromagnetism. The asymmetry around $x = 1/2$ is due to the asymmetry between electron and hole pockets. We note that this gradient term can be incorporated into equation (2) by completing the square, and replacing q_0 by:

$$q_{\text{eff}} = q_0 - \frac{d_2}{\xi_\rho^2}M^2 = \frac{1}{2} - x - \frac{d_2}{\xi_\rho^2}M^2 \quad (3)$$

The sign of d_2 is unknown *a priori* and here we choose it to be positive. Once this sign is fixed, however, this gradient coupling has profound consequences for the phase diagram. First, note that even if we are at commensurability ($x = 1/2$), if magnetism is present, then there is a tendency to incommensurate charge modulation. This reproduces the experiments of ref. 7 on $\text{La}_{1/2}\text{Ca}_{1/2}\text{MnO}_3$, where the onset of charge modulation is incommensurate, and accompanied by ferromagnetism—which is replaced by Néel

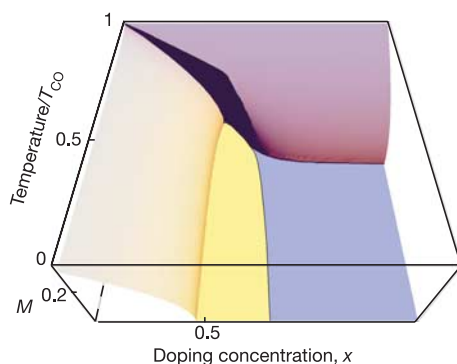


Figure 3 Magnetization in the doping–temperature plane within phase modulation approximation for a particular choice of parameters. This complements and clarifies the schematic view in Fig. 2. The low-temperature phases are non-magnetic for $x > 0.5$ (corresponding to antiferromagnetism in experiments) and weakly ferromagnetic for $x < 0.5$ (corresponding to the canted antiferromagnetism found in $\text{Pr}_{1-x}\text{Ca}_x\text{MnO}_3$; refs 17, 22). Above the Néel temperature there is a re-entrant magnetization that has been reported for $x \geq 0.5$ (ref. 21).

order at the transition to the commensurate phase. The incommensurate phase of $\text{Pr}_{1/2}\text{Ca}_{1/2}\text{MnO}_3$ (see Fig. 1a) is paramagnetic but accompanied by the onset of ferromagnetic spin fluctuations^{14,26}. The second feature of this term is that if $x < 1/2$, it is possible for coexisting magnetism to 'cancel' the chemical tendency to incommensurability, and we note that canted magnetism is generally reported^{17,22} in the underdoped manganites that show commensurate charge modulation. No such cancellation is possible for $x > 1/2$, and thus the dog-leg dependence of the wave vector on doping shown in the inset to Fig. 1a is indicated.

These features are reproduced by numerical minimization of the coupled free energy with appropriate values of the parameters (see Supplementary Information). Figure 2 shows the generic form of the phase diagram that can be obtained. In Figs 1b and 3 we show an explicit evaluation of the temperature- and doping-dependence of the (in)commensurability and magnetism respectively for parameters chosen to approximately reproduce the experimental regimes. The parameter range that can be used without altering the main features in Figs 1b, 2 and 3 is quite wide provided that charge modulation dominates over magnetic order.

Throughout much of the phase diagram shown in these figures, the order parameters M , ρ and $\nabla\phi$ are approximately uniform in space, at least for the parameters we have chosen. However, close to the commensurate–incommensurate transition, the phase modulation becomes non-uniform, and the phase gradient is built up by periodic discommensurations where the phase advances through $2\pi/n$. In such an inhomogeneous state, magnetism is naturally enhanced at the boundary, and the amplitude of the charge modulation is suppressed (see Fig. 4). Another situation where inhomogeneity is enforced is at a magnetic domain wall, where it can be energetically preferable to have a sharp wall stabilized by an insertion of local charge modulation. Such a phenomenon might be responsible for the anomalously large resistance reported for magnetic domain walls in LaCaMnO_3 (refs 27, 28).

When interpreting experimental results in the light of the theory presented here, a couple of issues must be kept in mind. First, the theory proposed is a mean field theory that cannot describe strong fluctuations because its solutions are uniformly ordered or disordered phases. However, experimentally there are some regimes,

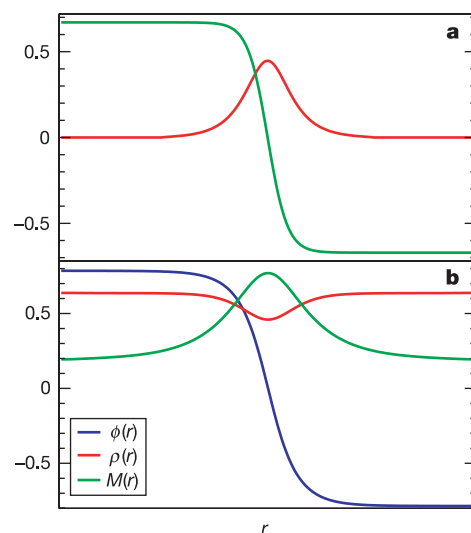


Figure 4 Example solutions of the total free energy when all the order parameters are allowed to change spatially. **a**, The magnetization (green) decays in the centre of a magnetic domain wall leading to the appearance of charge modulation (red). **b**, In a discommensuration, the phase of the charge modulation (blue) changes by $\pi/2$, the amplitude of the charge modulation (red) is suppressed, and the magnetization (green) is enhanced.

like the incommensurate phase of $\text{Pr}_{1/2}\text{Ca}_{1/2}\text{MnO}_3$, where strong ferromagnetic fluctuations have been measured^{14,26}, rather than the long-range magnetic order we would predict. Second, in order to keep the calculations tractable, the phase transitions at T_C and T_{CO} have been forced to be continuous. This explains why the re-entrant magnetism above T_1 shown in Fig. 2 appears only when $T_C > T_{CO}$. If this theory were generalized to include discontinuous phase transitions, this condition would relax. Another consequence of assuming continuous-phase transitions is that phase separation cannot be predicted. In real systems, phase separation is possible because strain^{3,9}, or disorder⁸, can make more-or-less localized phases dominate within a given region. Orbital ordering is described by a vector order parameter, and thus our simple model cannot address the complexity of different orbitally ordered phases that have been proposed¹².

The Ginzburg–Landau phenomenology we propose is capable of systematizing some puzzling data for manganites near $x = 1/2$, but of course the propensity for mixed and homogeneous phases is driven by the underlying physical parameters that make the energetic cost of spatial fluctuations low. This ‘electronic softness’ means that as well as spatially disordered ‘phase separation’, we find new ordered phases which are long-period arrangements of the two competing orders. Indeed, it may be that this potential for textured electronic phases is a hallmark of electronic oxides near the Mott transition²⁹, seen perhaps in the coexistence of density waves and superconductivity in the copper oxides³⁰. □

Received 29 July; accepted 17 December 2004; doi:10.1038/nature03300.

1. Tokura, Y. (ed.) *Colossal Magnetoresistance Oxides* (Gordon and Breach, New York, 2000).
2. Mathur, N. D. & Littlewood, P. B. Mesoscopic textures in manganites. *Phys. Today* **56**, 25–30 (2003).
3. Mathur, N. D. & Littlewood, P. B. The self-organised phases of manganites. *Sol. Stat. Commun.* **119**, 271–280 (2001).
4. Dagotto, E. *Nanoscale Phase Separation and Colossal Magnetoresistance* (Springer Series in Solid State Sciences Vol. 136, Springer, 2002).
5. Uehara, M., Mori, S., Chen, C. H. & Cheong, S.-W. Percolative phase separation underlies colossal magnetoresistance in mixed-valence manganites. *Nature* **399**, 560–563 (1999).
6. Loudon, J. C., Mathur, N. D. & Midgley, P. A. Charge-ordered ferromagnetic phase in $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$. *Nature* **420**, 797–800 (2002).
7. Chen, C. H. & Cheong, S.-W. Commensurate to incommensurate charge ordering and its real-space images in $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$. *Phys. Rev. Lett.* **76**, 4042–4045 (1996).
8. Burgu, J., Moreo, A. & Dagotto, E. Relevance of cooperative lattice effects and correlated disorder in phase separation theories for CMR manganites. *Phys. Rev. Lett.* **92**, 097202 (2004).
9. Ahn, K. H., Lookman, T. & Bishop, A. R. Strain-induced metal–insulator phase coexistence in perovskite manganites. *Nature* **428**, 401–404 (2004).
10. Goodenough, J. Theory of the role of covalence in the perovskite-type manganites $[\text{La}, \text{M(II)}]\text{MnO}_3$. *Phys. Rev.* **100**, 564–573 (1955).
11. Herrero-Martin, J., Garcia, J., Subias, G., Blasco, J. & Sánchez, M. C. Structural origin of dipole x-ray resonant scattering in the low-temperature phase of $\text{Nd}_{0.5}\text{Sr}_{0.5}\text{MnO}_3$. *Phys. Rev. B* **70**, 024408 (2004).
12. Brey, L. Continuous charge modulated diagonal phase in manganites. *Phys. Rev. Lett.* **92**, 127202 (2004).
13. Coey, M. Charge ordering in oxides. *Nature* **430**, 155–156 (2004).
14. Kajimoto, R., Yoshizawa, H., Tomioka, Y. & Tokura, Y. Commensurate–incommensurate transition in the melting process of orbital ordering in $\text{Pr}_{0.7}\text{Ca}_{0.3}\text{MnO}_3$: A neutron diffraction study. *Phys. Rev. B* **63**, 212407 (2001).
15. Zimmermann, M. v. et al. X-ray resonant scattering studies of orbital and charge ordering in $\text{Pr}_{1-x}\text{Ca}_x\text{MnO}_3$. *Phys. Rev. B* **64**, 195133 (2001).
16. Larochelle, S. et al. Nature of e_g electron order in $\text{La}_{1-x}\text{Sr}_{1+x}\text{MnO}_4$. *Phys. Rev. Lett.* **87**, 095502 (2001).
17. Jirak, Z., Krupicka, S., Simsa, Z., Dlouham, M. & Vratislav, S. Neutron diffraction study of $\text{Pr}_{1-x}\text{Ca}_x\text{MnO}_3$ perovskites. *J. Magn. Magn. Mater.* **53**, 153–166 (1985).
18. Chen, C. H., Cheong, S.-W. & Hwang, H. Y. Charge ordered stripes in $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$. *J. Appl. Phys.* **81**, 4326–4330 (1997).
19. Chen, C. H., Mori, S. & Cheong, S.-W. Anomalous melting transition of the charge ordered state in manganites. *Phys. Rev. Lett.* **83**, 4792–4795 (1999).
20. Loudon, J. C. et al. Weak charge–lattice coupling requires reinterpretation of stripes of charge order in $(\text{La}, \text{Ca})\text{MnO}_3$. Preprint at (<http://www.arXiv.org/cond-mat/0308581>) (2003).
21. Schiffer, P., Ramirez, A. P., Bao, W. & Cheong, S.-W. Low temperature magnetoresistance and the magnetic phase diagram of $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$. *Phys. Rev. Lett.* **75**, 3336–3339 (1995).
22. Yoshizawa, H., Kawano, H., Tomioka, Y. & Tokura, Y. Neutron diffraction study of the magnetic-field-induced metal–insulator transition in $\text{Pr}_{0.7}\text{Ca}_{0.3}\text{MnO}_3$. *Phys. Rev. B* **52**, R13145–R13148 (1995).
23. Tolédano, J. C. & Tolédano, P. *The Landau Theory of Phase Transitions* Ch. 5 (World Scientific, Singapore, 1987).
24. Murakami, S. & Nagaosa, N. Colossal magnetoresistance in manganites as multicritical phenomena. *Phys. Rev. Lett.* **90**, 197201 (2003).
25. Yamada, Y. & Takakura, T. Incommensurate orbital order and two-phase coexistence in doped manganites at metal–insulator phase boundary. *J. Phys. Soc. Jpn* **71**, 2480–2484 (2002).
26. Kajimoto, R. et al. Anomalous ferromagnetic spin fluctuations in an antiferromagnetic insulator $\text{Pr}_{1-x}\text{Ca}_x\text{MnO}_3$. *Phys. Rev. B* **58**, R11837–R11840 (1998).

27. Mathur, N. D. et al. Resistance of a domain wall in $\text{La}_{0.7}\text{Ca}_{0.3}\text{MnO}_3$. *J. Appl. Phys.* **86**, 6287–6290 (1999).
28. Rychowski, M. S. & Joynt, R. Electronic inhomogeneity at magnetic domain walls in strongly-correlated systems. *Europhys. Lett.* **67**, 287–293 (2004).
29. Kivelson, S. A., Fradkin, E. & Emery, V. J. Electronic liquid–crystal phases of a doped Mott insulator. *Nature* **393**, 550–553 (1998).
30. McElroy, K. et al. Homogenous nodal superconductivity coexisting with inhomogeneous charge order in strongly underdoped Bi-2212 . Preprint at (<http://www.arXiv.org/cond-mat/0404005>) (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank L. Brey and N. Mathur for discussions. P.B.L. thanks the National High Magnetic Field Laboratory of the Los Alamos National Laboratory for hospitality. M.J.C. acknowledges Churchill College, University of Cambridge, for the award of a JRE. This work was supported by the EPSRC and through the EPSRC Magnetic Oxide Network.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.C.M. (gcm24@cam.ac.uk).

Synthesis of the H-cluster framework of iron-only hydrogenase

Cédric Tard¹, Xiaoming Liu¹, Saad K. Ibrahim¹, Maurizio Bruschi³, Luca De Gioia³, Siân C. Davies¹, Xin Yang⁴, Lai-Sheng Wang⁴, Gary Sawers² & Christopher J. Pickett¹

¹Department of Biological Chemistry, ²Department of Molecular Microbiology, John Innes Centre, Norwich NR4 7UH, UK

³Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza della Scienza 2, 20126-Milan, Italy

⁴Department of Physics, Washington State University, 2710 University Drive, Richland, and WR Wiley Environmental Science Laboratory and Chemical Sciences Division, Pacific Northwest National Laboratory, PO Box 999, MS K8-88, Richland, Washington 99352, USA

The metal–sulphur active sites of hydrogenases catalyse hydrogen evolution or uptake at rapid rates. Understanding the structure and function of these active sites—through mechanistic studies of hydrogenases^{1–4}, synthetic assemblies^{5–12} and *in silico* models^{13–15}—will help guide the design of new materials for hydrogen production or uptake¹⁶. Here we report the assembly of the iron–sulphur framework of the active site of iron-only hydrogenase (the H-cluster), and show that it functions as an electrocatalyst for proton reduction. Through linking of a di-iron subsite to a {4Fe4S} cluster, we achieve the first synthesis of a metallosulphur cluster core involved in small-molecule catalysis. In addition to advancing our understanding of the natural biological system, the availability of an active, free-standing analogue of the H-cluster may enable us to develop useful electrocatalytic materials for application in, for example, reversible hydrogen fuel cells. (Platinum is currently the preferred electrocatalyst for such applications, but is expensive, limited in availability and, in the long term, unsustainable¹⁷.)

The crystallographic characterization of Fe-only hydrogenases^{1,2} has revealed a striking resemblance of the di-iron subsite of the H-cluster to known $[\text{Fe}_2(\mu\text{-SR})_2(\text{CO})_6]$ (R = organic group) complexes. This type of assembly, first discovered¹⁸ more than 70 years ago, opened the way for the synthesis of {2Fe2S}- and {2Fe3S}-complexes with key structural and/or spectroscopic features of this biologically unprecedented—low-valent, carbon monoxide- and cyanide-coordinated—di-iron unit (Fig. 1). A major challenge is now to build a free-standing analogue of the entire H-cluster, as this offers the prospects of understanding the interplay of the conjoined di-iron and cubane units that form the enzymic catalytic machinery

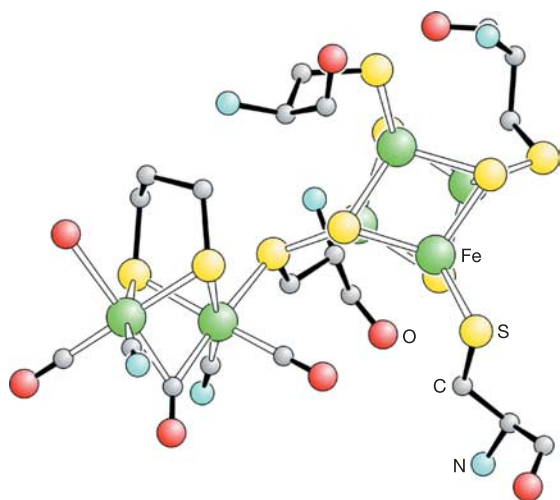


Figure 1 Composite structure of the H-cluster. This was constructed from the crystal structures of Fe-only hydrogenase isolated from *Desulfovibrio desulfuricans* (Protein Data Bank Code 1HFE)² and *Clostridium pasteurianum* (Protein Data Bank Code 1FEH)¹, and FTIR data from *Desulfovibrio vulgaris*. The apical group on the sub-site ligand may possibly be an NH, but this remains crystallographically and analytically unresolved.

and of the development of new catalytic materials.

The dithiolate ligand **A**, which possesses an appended thioester group, was synthesized as outlined in Fig. 2. **A** reacts with $\text{Fe}_3(\text{CO})_{12}$ to give the thioester activated di-iron subsite analogue, **B**. Unlike dithiolate thioether ligands with the same tripodal carbon backbone¹⁹, the thioester sulphur does not displace CO and coordinate to an Fe atom of the di-iron unit, as shown in the X-ray crystallographic structure (Supplementary Fig. 1). Undoubtedly this is because the electron-withdrawing acyl group lowers the nucleophilicity of the S atom to which it is attached.

The cubane cluster $[\text{Fe}_4\text{S}_4(\text{L})(\text{SEt})]^{2-}$ ($\text{L} = 1,3,5$ -tris (4,6-dimethyl-3-mercaptophenylthio)-2,4,6-tris (*p*-tolyl-thio)benzene), denoted **C**, which has three of the cubane iron atoms blocked by the chelating ligand, was synthesized by an established procedure²⁰. **C** reacts cleanly with one equivalent of **B** over a 12 h period at room temperature to give, after work-up, analytically pure $[\text{Fe}_4\text{S}_4(\text{L})_3\{\text{Fe}_2(\text{CH}_3\text{C}(\text{CH}_2\text{S})_3)(\text{CO})_5\}][\text{NBu}_4]_2$, **D** (Fig. 2).

Negative ion electrospray mass spectrometry (ESMS; MeCN) on **D** showed peaks centred at mass/charge (m/z) = 856.9 (100%) for the dianion and at m/z = 1,957.1 (10%) for $\{[\text{Fe}_4\text{S}_4(\text{L})_3\{\text{Fe}_2(\text{CH}_3\text{C}(\text{CH}_2\text{S})_3)(\text{CO})_5\}]^{2-}[\text{NBu}_4]^+\}$, with respective isotopic distribution patterns consistent with doubly and singly charged species. Fourier transform infrared (FTIR) $\nu(\text{CO})$ bands for **D** at 2,035 (medium, m), 1,970 (strong, s) and 1,912 (weak, w) cm^{-1} are very similar to those of $[\text{Fe}_2(\text{CH}_3\text{C}(\text{CH}_2\text{S})_2(\text{CH}_2\text{SCH}_3))(\text{CO})_5]$ (2,049 (m), 1,983 (s) and 1,927 (w) cm^{-1}) in which the thioether ligand is known to be coordinated to one Fe atom¹⁹. Thus the reaction between **B** and **C** proceeds beyond elimination of MeCOSet—carbon monoxide is lost from the proximal Fe atom of the di-iron unit, and results in the formation of a $\{\text{Fe}_{\text{cubane}}(\mu\text{-SR})\text{Fe}_{\text{subsite}}\}$ linkage as found in the H-cluster (Fig. 3). This is supported by the solid state Mössbauer spectrum of **D** at 80 K, which exhibits four overlapping quadrupole split doublets with isomer shift (i.s.) and quadrupole splitting (q.s.) parameters (in mm s^{-1}) consistent with four differentiated iron sites; two associated with the site-differentiated cubane (i.s. = 0.47, q.s. = 1.14; i.s. = 0.45, q.s. = 0.87) and two with the di-iron subsite (i.s. = 0.04, q.s. = 1.00 (Fe distal to cubane); i.s. = 0.04, q.s. = 0.50 (Fe proximal to cubane))¹⁹.

Full geometry optimization of an *in silico* model of **D**, $[\text{Fe}_4\text{S}_4(\text{SCH}_3)_3\{\text{Fe}_2(\text{CH}_3\text{C}(\text{CH}_2\text{S})_3)(\text{CO})_5\}]^{2-}$, was carried out in the density functional theory (DFT) framework using the BP86 pure functional^{21,22} and an all-electron valence triple- ζ basis set with polarization functions on all atoms (TZVP)²³. This resulted in a structure fully consistent with that proposed on the basis of experimental data for **D** (Fig. 3). In particular, one of the thiolate groups bridges almost symmetrically the $\{4\text{Fe}4\text{S}\}$ and the $\{2\text{Fe}2\text{S}\}$ units, with an Fe–Fe distance (2.6 Å) in the binuclear cluster that is indicative of a metal–metal bond. Analysis of the electronic properties of the *in silico* structure reveals that the redox state of the binuclear moiety can be described as $\text{Fe}(\text{I})\text{Fe}(\text{I})$.

The $\{\text{Fe}_{\text{cubane}}(\mu\text{-SR})\text{Fe}_{\text{subsite}}\}$ linkage provides substantial electronic communication between the di-iron centre and the $\{4\text{Fe}4\text{S}\}$ cubane centre, as is evident from the following spectroscopic and electrochemical measurements. Formally replacing the Me of the thioether ligand in $[\text{Fe}_2(\text{CH}_3\text{C}(\text{CH}_2\text{S})_2(\text{CH}_2\text{SCH}_3))(\text{CO})_5]$ by the cubane dianion shifts all the $\nu(\text{CO})$ frequencies of the appended subsite by $\sim 15 \text{ cm}^{-1}$ to lower values: the $\{\text{Fe}_4\text{S}_4(\text{L})\}^{2-}$ -core is a 'better' donor group than is methyl. The inductive influence of the cubane in lowering $\nu(\text{CO})$ is reciprocated in the shift of the redox

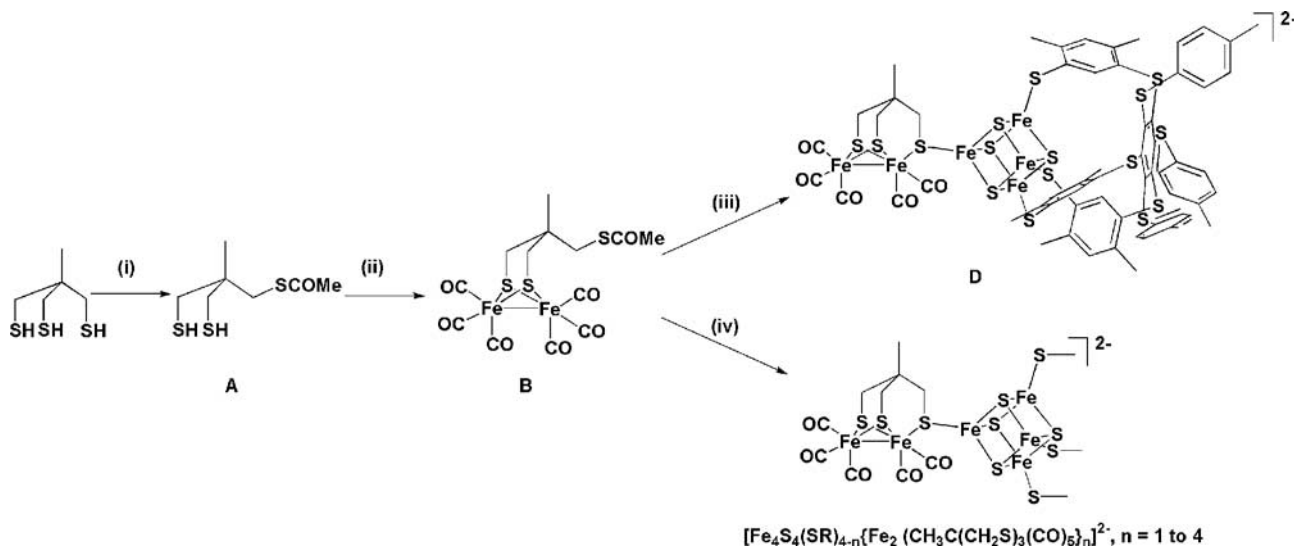


Figure 2 Synthetic pathways for assembly of the H-cluster model and related subsite-cluster materials. Reaction steps are as follows: (i) acetic anhydride, NaHCO_3 , diethyl ether, (ii) $\text{Fe}_3(\text{CO})_{12}$, toluene, (iii) **C**, MeCN, (iv) $[\text{Fe}_4\text{S}_4(\text{SEt})_4][\text{NBu}_4]_2$, MeCN.

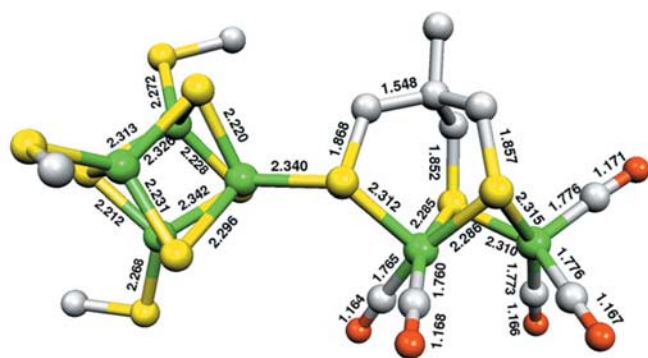


Figure 3 Structure of $[\text{Fe}_4\text{S}_4(\text{SCH}_3)_3\{\text{Fe}_2(\text{CH}_3\text{C}(\text{CH}_2\text{S})_3(\text{CO})_5)\}]^{2-}$, derived from DFT calculations. Numbers show bond lengths in Å.

potential for the reduction of the cubane core to more positive potentials. The primary reversible one-electron reduction of **D** ($0.1 \text{ M } [\text{NBu}_4][\text{BF}_4]\text{-MeCN}$, cyclic voltammetry, sweep rate 100 mV s^{-1}) occurs at $E_{1/2} = -0.86 \text{ V}$ versus (Ag/AgCl, CH_2Cl_2 , $0.45 \text{ M } [\text{NBu}_4][\text{BF}_4]$, $0.05 \text{ M } [\text{NBu}_4]\text{Cl}$); that for the parent cluster $[\text{Fe}_4\text{S}_4(\text{L})(\text{SEt})]^{2-}$ is at $E_{1/2} = -0.98 \text{ V}$ under identical conditions. The $\Delta E_{1/2}$ for the formal replacement of EtS by $\{\text{Fe}_2(\text{CH}_3\text{C}(\text{CH}_2\text{S})_3(\text{CO})_5)\}$ of $+120 \text{ mV}$ shows that the energy of the cubane lowest unoccupied molecular orbital (LUMO) is substantially lowered.

Gas-phase photoelectron spectroscopy (PES)^{24,25} of isolated $[\text{Fe}_4\text{S}_4(\text{SEt})_3\{\text{Fe}_2(\text{CH}_3\text{C}(\text{CH}_2\text{S})_3(\text{CO})_5)\}]^{2-}$ is concordant with this. The PES of this anion at an excitation wavelength of 355 nm revealed two low-binding-energy spectral features that are very similar to the first two bands of $[\text{Fe}_4\text{S}_4(\text{SEt})_4]^{2-}$ but shifted by about 0.6 eV to higher energy (adiabatic electron detachment energy increases from 0.29 to 0.89 eV (ref. 26); Supplementary Fig. 2). Thus on replacing EtS by the subsite unit the highest-energy orbitals remain localized on the cubane core but are substantially stabilized. At 266 and 193 nm , higher-energy spectral features associated with the attached subsite are also observed that are absent in the parent dianion, $[\text{Fe}_4\text{S}_4(\text{SEt})_4]^{2-}$.

Complete substitution of the $\{4\text{Fe}4\text{S}\}$ -core by di-iron units is possible (Fig. 2, step (iv)). The reaction of $[\text{Fe}_4\text{S}_4(\text{SEt})_4][\text{NBu}_4]_2$ (MeCN, room temperature, 12 h) with an excess of **B** followed by work-up and ESMS shows formation of a dianion with m/z centred at $1,009.6$ (100%) and an isotopic pattern corresponding to the dianion $[\text{Fe}_4\text{S}_4\{\text{Fe}_2(\text{CH}_3\text{C}(\text{CH}_2\text{S})_3(\text{CO})_5\}_4]^{2-}$, **E**. Cyclic

voltammetry of **E** (vitreous carbon electrode, 100 mV s^{-1} , room temperature, $0.1 \text{ M } [\text{NBu}_4][\text{BF}_4]\text{-CH}_2\text{Cl}_2$) shows a primary reversible one-electron reduction couple at -0.65 V versus (Ag/AgCl, CH_2Cl_2 , $0.45 \text{ M } [\text{NBu}_4][\text{BF}_4]$, $0.05 \text{ M } [\text{NBu}_4]\text{Cl}$) corresponding to the reduction of the $\{4\text{Fe}4\text{S}\}^{2+}$ core, and a multi-electron reduction at -1.58 V which encompasses the irreversible reduction of the subsite units. The primary reversible reduction of $[\text{Fe}_4\text{S}_4(\text{SEt})_4]^{2-}$ occurs at -1.15 V under the same conditions, attesting to the strong electron-withdrawing properties of the $\mu\text{-S}$ ligated di-iron subsites. The ΔE of $+500 \text{ mV}$ for replacement of the four EtS ligands by four subsite ligands (that is, an average of 125 mV per substitution) fits very well with the 120 mV shift observed on replacing the single EtS with the di-iron subsite to give **D**.

Beyond inductive effects transmitted through the $\{2\text{Fe}2\text{S}\}(\mu\text{-SCH}_2\text{R})\{4\text{Fe}4\text{S}\}$ framework, there is evidence for an interplay of redox states that transcends the behaviour of the isolated component di-iron and the cubane centre. Repetitive cyclic voltammetry of **D** over 10 cycles shows the build-up of a reversible system at -0.96 V versus (Ag/AgCl, CH_2Cl_2 , $0.45 \text{ M } [\text{NBu}_4][\text{BF}_4]$, $0.05 \text{ M } [\text{NBu}_4]\text{Cl}$) (Fig. 4a and Supplementary Fig. 3 (process II)). The voltammogram can be reasonably simulated by a mechanism that involves intramolecular electron transfer between the reduced cluster and subsite in concert with reversible opening of the $\mu\text{-SCH}_2$ bridge. This leaves the cubane linked to the subsite by an alkylthiolate in a $\{4\text{Fe}4\text{S}\}^{2+}$ state and therefore susceptible to further reduction. This would be expected to take place at a potential close to that observed for the alkylthiolate ligated $[\text{Fe}_4\text{S}_4(\text{L})(\text{SEt})]^{2-}$, as is observed (Supplementary Fig. 4 and Supplementary Text).

Studies^{11,27} have shown that simple subsite models are capable of electrocatalysing proton reduction and H/D exchange reactions, albeit at low reduction potentials^{12,28}. Figure 4b shows that the H-cluster analogue **D** electrocatalyses proton reduction at a diffusion-controlled rate. The current–potential curve for the reduction of protons in the presence of the catalyst ($E_p = -1.13 \text{ V}$) is displaced by $\sim 200 \text{ mV}$ positive from that measured in the absence of the catalyst ($E_p = -1.33 \text{ V}$ versus (Ag/AgCl, CH_2Cl_2 , $0.45 \text{ M } [\text{NBu}_4][\text{BF}_4]$, $0.05 \text{ M } [\text{NBu}_4]\text{Cl}$)).

The $\mu\text{-S}$ cysteinyl bridge identified in the enzyme structures^{1,2} must provide substantial electronic communication between the subsite and the $\{4\text{Fe}4\text{S}\}$ -cubane, because we have shown for **D** that these units do not behave as insulated redox entities. The Mössbauer parameters for **D** corrected for the second order Doppler effect to 4.2 K (site differentiated cubane: i.s. = 0.50 , 0.48 , q.s. = 0.87 , 1.14 ; di-iron subsite: i.s. = 0.07 , 0.07 q.s. = 0.50 , 1.00 mm s^{-1}) are similar to those attributed to these components in the reduced state of *Clostridium pasteurianum* hydrogenase II (site differentiated

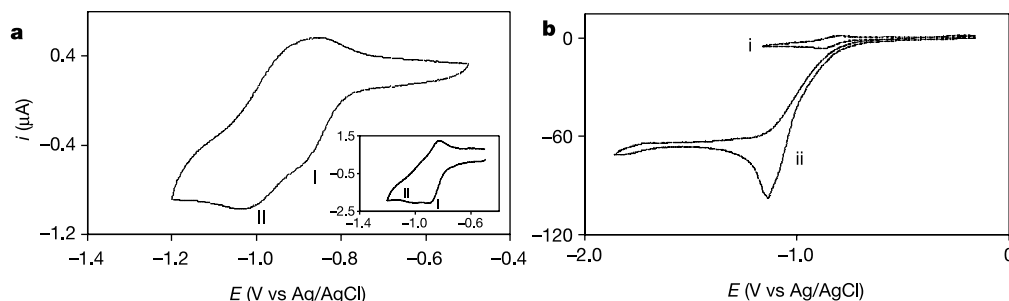


Figure 4 The electrochemical behaviour of the synthetic H-cluster model. **a**, Cyclic voltammogram of **D** ($0.1 \text{ M } [\text{NBu}_4][\text{BF}_4]\text{-MeCN}$, 30 mV s^{-1} , vitreous carbon), showing interconverting redox processes I and II. $E_{1/2}$ for II occurs at a potential close to that for **C**, indicative of the rearrangement of the $\mu\text{-S}$ bonded subsite to the terminal alkylthiolate. Inset, first-scan response at the faster scan rate of 0.1 V s^{-1} but otherwise under the same conditions; note suppression of the interconversion. **b**, Cyclic voltammogram of **D**

(1.5 mM), showing the electrocatalytic response in the absence (trace i) and in the presence (trace ii) of 4,6-dimethyl pyridinium cation as a source of protons (15 mM) at a normalized scan rate of 0.1 V s^{-1} . The peak current for proton reduction in the absence of the catalyst is within 5% of that in its presence, but is shifted about 200 mV to a more negative potential.

cubane: i.s. = 0.47, 0.47, q.s. = 0.85, 1.35; di-iron subsite: i.s. = 0.08, q.s. = 0.87 mm s⁻¹) measured at 4.2 K. This is concordant with the generally accepted re-assignment of the electronic structure of the six-iron core of the reduced biological cluster as [Fe(I).Fe(I)]_{subsite}-[4Fe4S]²⁺_{cubane}]²⁹, the established redox configuration of our synthetic cluster. In the synthetic system, we have seen that the reduction of the cubane unit by one electron to the {4Fe4S}⁺ level is easier than is reduction of the {Fe(I).Fe(I)}_{subsite}, which is coordinatively saturated with a closed-shell (36-electron) configuration. A corresponding state of the H-cluster, in which the cubane unit is reduced to the {4Fe4S}⁺ level, has yet to be detected, although neighbouring [4Fe4S] relay centres in *C. pasteurianum* hydrogenase II can be reduced to this level²⁹. This raises the question as to whether the {4Fe4S}⁺ level of the H-cluster is physiologically accessed during turnover. A vacant or (weakly) water-coordinated site has been identified crystallographically in the enzyme at the distal iron atom in the resting state of the enzyme. It is possible that protonation at this site lowers the energy of the Fe(I)-Fe(I) subsite unit sufficiently to enable its reduction by the anchored cubane operating at the {4Fe4S}²⁺ level. That D is capable of electrocatalysing proton reduction may be similarly linked to the formation of a vacant site, in this case by the opening of the μ -S bridge on reduction.

The artificial H-clusters reported here should enhance our understanding of the intimate chemistry of the natural process, and lead to systems with low overpotentials for hydrogen uptake/evolution^{11,28}. Given that redox-active {4Fe4S}²⁺-centres can be incorporated at high concentration into cysteine functionalized electropolymers³⁰, we can envisage their modification, using the chemistry we have described, thereby providing a route to advanced electrode materials. □

Received 14 October; accepted 14 December 2004; doi:10.1038/nature03298.

- Peters, J. W., Lanzilotta, W. N., Lemon, B. J. & Seefeldt, L. C. X-ray crystal structure of the Fe-only hydrogenase (Cpl) from *Clostridium pasteurianum* to 1.8 angstrom resolution. *Science* **282**, 1853–1858 (1998).
- Nicolet, Y., Piras, C., Legrand, P., Hatchikian, C. E. & Fontecilla-Camps, J. C. *Desulfovibrio desulfuricans* iron hydrogenase: the structure shows unusual coordination to an active site Fe binuclear center. *Struct. Fold. Des.* **7**, 13–23 (1999).
- Nicolet, Y., Lemon, B. J., Fontecilla-Camps, J. C. & Peters, J. W. A novel FeS cluster in Fe-only hydrogenases. *Trends Biochem. Sci.* **25**, 138–143 (2000).
- Armstrong, F. A. Hydrogenases: active site puzzles and progress. *Curr. Opin. Chem. Biol.* **8**, 133–140 (2004).
- Evans, D. J. & Pickett, C. J. Chemistry and the hydrogenases. *Chem. Soc. Rev.* **32**, 268–275 (2003).
- Lyon, E. J., Georgakaki, I. P., Reibenspies, J. H. & Darensbourg, M. Y. Carbon monoxide and cyanide ligands in a classical organometallic complex model for Fe-only hydrogenase. *Angew. Chem. Int. Edn* **38**, 3178–3180 (1999).
- Schmidt, M., Contakes, S. M. & Rauchfuss, T. B. First generation analogues of the binuclear site in the Fe-only hydrogenases: Fe₂(μ -SR)₂(CO)₄(CN)₂⁻. *J. Am. Chem. Soc.* **121**, 9736–9737 (1999).
- Le Cloirec, A. et al. A di-iron dithiolate possessing structural elements of the carbonyl/cyanide sub-site of the H-centre of Fe-only hydrogenase. *Chem. Commun.* 2285–2286 (1999).
- Razavet, M. et al. Transient FTIR spectroelectrochemical and stopped-flow detection of a mixed valence [Fe(I)-Fe(II)] bridging carbonyl intermediate with structural elements and spectroscopic characteristics of the di-iron sub-site of all-iron hydrogenase. *Chem. Commun.* 700–701 (2002).
- George, S. J., Cui, Z., Razavet, M. & Pickett, C. J. The di-iron subsite of all-iron hydrogenase: Mechanism of cyanation of a synthetic [2Fe3S] – carbonyl assembly. *Chem. Eur. J.* **8**, 4037–4046 (2002).
- Gloaguen, E., Lawrence, J. D., Rauchfuss, T. B., Benard, M. & Rohmer, M. M. Bimetallic carbonyl thiolates as functional models for Fe-only hydrogenases. *Inorg. Chem.* **41**, 6573–6582 (2002).
- Ott, S., Kritikos, M., Akermark, B., Sun, L. C. & Lomoth, R. A biomimetic pathway for hydrogen evolution from a model of the iron hydrogenase active site. *Angew. Chem. Int. Edn* **43**, 1006–1009 (2004).
- Cao, Z. X. & Hall, M. B. Modeling the active sites in metalloenzymes. 3. Density functional calculations on models for Fe-hydrogenase: Structures and vibrational frequencies of the observed redox forms and the reaction mechanism at the diiron active center. *J. Am. Chem. Soc.* **123**, 3734–3742 (2001).
- Liu, Z. P. & Hu, P. A density functional theory study on the active center of Fe-only hydrogenase: Characterization and electronic structure of the redox states. *J. Am. Chem. Soc.* **124**, 5175–5182 (2002).
- Bruschi, M., Fantucci, P. & De Gioia, L. Density functional theory investigation of the active site of Fe-hydrogenases. Systematic study of the effects of redox state and ligands hardness on structural and electronic properties of complexes related to the [2Fe]_H subcluster. *Inorg. Chem.* **43**, 3733–3741 (2004).
- Basic Research Needs for the Hydrogen Economy (Report of the Basic Energy Sciences Workshop on Hydrogen Production, Storage, and Use, 13–15 May 2003, Office of Science, US Department of Energy); available at (http://www.eere.energy.gov/hydrogenandfuelcells/pdfs/bes_project.pdf).

- Platinum and Hydrogen for Fuel Cell Vehicles (UK Department for Transport, September 2003); available at (http://www.dft.gov.uk/stellent/groups/dft_roads/documents/page/dft_roads_024056.hcsp).
- Reihlen, H., Gruhl, A. & Hessling, G. Über den photochemischen und oxydativen Abbau von Carbonsäuren. *Liebigs Ann. Chem.* **472**, 268–287 (1929).
- Razavet, M. et al. All-iron hydrogenase: synthesis, structure and properties of [2Fe3S]-assemblies related to the di-iron sub-site of the H-cluster. *Dalton Trans.* 586–595 (2003).
- Stack, T. D. P. & Holm, R. H. Subsite-specific functionalization of the [4Fe-4S]²⁺ analog of iron sulfur protein clusters. *J. Am. Chem. Soc.* **109**, 2546–2547 (1987).
- Perdew, J. P. Density-functional approximation for the correlation-energy of the inhomogeneous electron-gas. *Phys. Rev. B* **33**, 8822–8824 (1986).
- Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic-behavior. *Phys. Rev. A* **38**, 3098–3100 (1988).
- Schäfer, A., Huber, C. & Ahlrichs, R. Fully optimized contracted Gaussian-basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **100**, 5829–5835 (1994).
- Wang, L. S., Ding, C. F., Wang, X. B. & Barlow, S. E. Photodetachment photoelectron spectroscopy of multiply charged anions using electrospray ionization. *Rev. Sci. Instrum.* **70**, 1957–1966 (1999).
- Yang, X., Razavet, M., Wang, X. B., Pickett, C. J. & Wang, L. S. Probing the electronic structure of the di-iron subsite of Fe-hydrogenase: A photoelectron spectroscopic study of Fe(I)-Fe(I) model complexes. *J. Phys. Chem. A* **107**, 4612–4618 (2003).
- Wang, X. B. et al. Probing the intrinsic electronic structure of the cubane [4Fe4S] cluster: Nature's favorite cluster for electron transfer and storage. *J. Am. Chem. Soc.* **125**, 14072–14081 (2003).
- Zhao, X. et al. Catalysis of H₂/D₂ scrambling and other H/D exchange processes by [Fe]-hydrogenase model complexes. *Inorg. Chem.* **41**, 3917–3928 (2002).
- Tard, C., Liu, X., Hughes, D. L. & Pickett, C. J. A novel {Fe^I-Fe^{II}-Fe^{II}-Fe^I} iron thiolate carbonyl assembly which electrocatalyses hydrogen evolution. *Chem. Commun.* 133–135 (2005).
- Popescu, C. V. & Munck, E. Electronic structure of the H cluster in Fe-hydrogenases. *J. Am. Chem. Soc.* **121**, 7877–7884 (1999).
- Pickett, C. J. & Ryder, K. S. Bioinorganic reaction centers on electrodes — Modified electrodes possessing amino-acid, peptide and ferredoxin-type groups on a poly(pyrrole) backbone. *J. Chem. Soc. Dalton Trans.* 2181–2189 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by the BBSRC and the John Innes Foundation. We thank D. J. Evans and J. E. Barclay for Mössbauer spectroscopy and discussions; L. Hill for ESMs; and R. R. Eady and R. A. Dixon for discussions and comments on the manuscript. Research done at Washington State University was supported by NIH and performed at EMSL, a national scientific user facility sponsored by the US DOE's Office of Biological and Environmental Research and located at PNNL, which is operated for DOE by Battelle.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.J.P. (chris.pickett@bbsrc.ac.uk). Crystal structure data for B have been deposited at the Cambridge Crystallographic Data Centre, and allocated the deposition number CCDC 256735.

Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data

Anders Moberg¹, Dmitry M. Sonechkin², Karin Holmgren³, Nina M. Datsenko² & Wibjörn Karlén³

¹Department of Meteorology, Stockholm University, SE-106 91 Stockholm, Sweden

²Dynamical-Stochastic Laboratory, Hydrometeorological Research Centre of Russia, Bolshoy Predtechensky Lane 11/13, Moscow 123 242, Russia

³Department of Physical Geography and Quaternary Geology, Stockholm University, SE-106 91 Stockholm, Sweden

A number of reconstructions of millennial-scale climate variability have been carried out in order to understand patterns of natural climate variability, on decade to century timescales, and the role of anthropogenic forcing^{1–8}. These reconstructions have mainly used tree-ring data and other data sets of annual to decadal resolution. Lake and ocean sediments have a lower time resolution, but provide climate information at multicentennial timescales that may not be captured by tree-ring data^{9,10}. Here we reconstruct Northern Hemisphere temperatures for the past 2,000 years by combining low-resolution proxies with

tree-ring data, using a wavelet transform technique¹¹ to achieve timescale-dependent processing of the data. Our reconstruction shows larger multicentennial variability than most previous multi-proxy reconstructions^{1–4,7}, but agrees well with temperatures reconstructed from borehole measurements¹² and with temperatures obtained with a general circulation model^{13,14}. According to our reconstruction, high temperatures—similar to those observed in the twentieth century before 1990—occurred around AD 1000 to 1100, and minimum temperatures that are about 0.7 K below the average of 1961–90 occurred around AD 1600. This large natural variability in the past suggests an important role of natural multicentennial variability that is likely to continue.

Global temperatures are currently rising, and future scenarios suggest large changes in precipitation patterns and temperatures¹⁵. Reconstructing past climate is essential for enhanced understanding of climate variability, and provides necessary background knowledge for improving predictions of future changes. From multi-proxy combinations^{1–4,7} of climate proxy data (for example, from tree rings, corals and ice cores), or from reconstructions based solely on tree-ring data^{5–6,16}, past surface temperatures have been inferred by calibration to instrumental temperature data using statistical relationships. Past surface temperature changes have also been inferred from temperature profiles along deep terrestrial boreholes¹², where physical laws rather than statistical relations are used to estimate temperature trends. Borehole measurements¹² and some reconstructions based on tree-ring data^{6,16} indicate a larger warming trend in the past 500 yr than do multi-proxy reconstructions^{1–4,7}.

Although differences in the amplitude of centennial temperature variability have been discussed in the literature^{8,15}, the picture with relatively small variability (shown by multi-proxy reconstructions^{1–4,7}) is arguably best known by a wider audience. One reason for this is the prominent role that the multi-proxy reconstruction by Mann *et al.*^{1,2} had in the latest IPCC report¹⁵ and in public media. Recent findings¹⁴, however, suggest that considerable underestimation of centennial Northern Hemisphere temperature variability may result when regression-based methods, like those used by Mann *et al.*^{1,2}, are applied to noisy proxy data with insufficient spatial representativity. It is thus important to explore other techniques and other data types, and also to use data from more regions than previously examined.

A view has been expressed that only tree-ring and other high-resolution data are useful for quantitative large-scale temperature reconstructions^{3,8}. Tree-ring data, however, have a well-documented difficulty in reliably reproducing multicentennial temperature variability¹⁷. Special standardization techniques for extracting multicentennial information exist^{6,16}, but they do not ensure an accurate reconstruction¹⁰. Natural archives with lower temporal resolution (often various sediments) have potential for providing climate information at multicentennial timescales that may not be captured by tree-ring data⁹. Low-resolution data, on the other hand, can sometimes have dating errors of up to more than 100 yr (ref. 9). Such errors are intolerable at interannual to multi-decadal timescales⁸, but if these data are used only for longer timescales the errors become relatively small in relation to the timescales of interest.

Our aim is to combine low-frequency climate information (contained in low-resolution proxy data) with high-frequency information (from tree-ring data) in order to derive a 2,000-yr-long Northern Hemisphere temperature reconstruction in which we avoid using each proxy type at timescales where it is most unreliable, but instead use it only where it has its greatest advantages. To achieve this, we developed a method that allows a timescale-dependent data decomposition by application of a wavelet transform¹¹ technique.

The number of available 2,000-yr-long local-to-regional scale

temperature proxy series is very limited⁸. We found seven tree-ring series and eleven low-resolution proxy series. For the period before AD 1400, this number is similar to previous reconstructions^{1–7}. Among the eleven low-resolution series were nine already calibrated to local/regional temperatures for different Northern Hemisphere regions, whereas two were only available in original proxy data units. The data set includes climatic information from the North American and Eurasian continents and from the Atlantic and Indian oceans (Fig. 1, Table 1; see Supplementary Information for further details and references).

Various averages (see Methods) of the nine already calibrated low-resolution series (Table 1, site numbers 2–10) depict, when low-pass filtered to highlight multicentennial variability (Fig. 2a), a temperature maximum in the ninth or tenth centuries and a minimum in the sixteenth century, with a total range of 0.6 to 0.9 K. This is a substantially larger multicentennial temperature range compared to previous multi-proxy reconstructions^{1–4,7}, exemplified in Fig. 2a with the Northern Hemisphere reconstruction by Mann and Jones⁷. The conflicting views provided by the low-resolution series and this reconstruction illustrate a clear need for further research to determine the size of natural low-frequency climate variability.

Simple averages of temperature proxy series, such as those shown in Fig. 2a, can yield adequate estimates of Northern Hemisphere century-scale mean-temperature anomalies¹⁴. But to obtain a reconstruction covering a complete range of timescales we created a wavelet transform¹¹, supposed to be approximately representative for the Northern Hemisphere (Fig. 2e), such that only information from the tree-ring records contributes to timescales less than 80 yr and all eleven low-resolution proxies contribute only to the longer timescales (see Methods for details). From the wavelet transform, which was calculated using standardized (that is, unitless) proxy series, a dimensionless reconstruction was obtained by calculating the inverse wavelet transform¹¹. Given the small number of tree-ring series, their contribution is best thought of as an approximation of the statistical character of variability at <80-yr scales. A sample size of eleven low-resolution series is reasonable for reconstructing >80-yr variability, given that the number of spatial degrees of freedom on the entire globe is about five at centennial time-

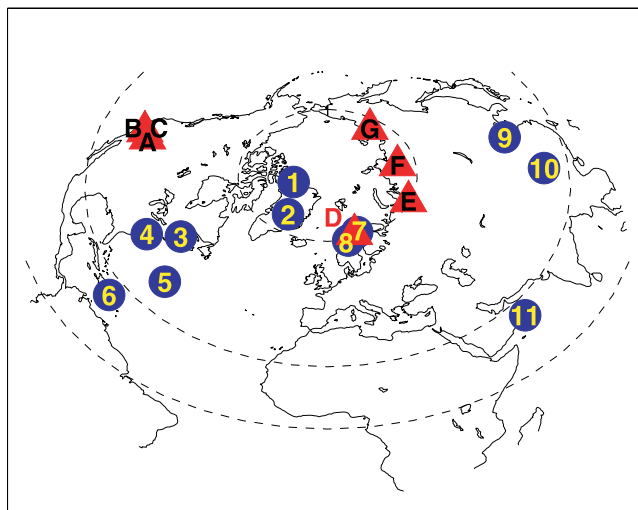


Figure 1 Locations of proxy data sites. Shown are the locations of eleven low-resolution proxy records (blue circles, numerals 1–11, see Table 1 for site details) and seven tree-ring series (red triangles, letters A–G). Dashed latitude lines indicate the Equator, the Tropic of Cancer and the Arctic Circle. See Supplementary Information for further details and references.

scales¹⁸. The contribution from each low-resolution proxy series to the total variance at >80-yr scales in the Northern Hemisphere reconstruction is given in Table 1.

To calibrate the reconstruction, its mean value and variance were adjusted to agree with the instrumental record of Northern Hemisphere annual mean temperatures¹⁹ in the overlapping period AD 1856–1979 (Fig. 2b). This technique avoids the problem with underestimation of low-frequency variability associated with regression-based calibration methods¹⁴. Jack-knifed estimates (light-blue curves in Fig. 2b) of the low-frequency component (>80-yr scales, dark blue), where each of the eleven low-resolution proxies are excluded one at a time, are all very similar, demonstrating that the reconstruction is robust to the omission of any single low-resolution proxy series.

The reconstruction depicts two warm peaks around AD 1000 and 1100 and pronounced coolness in the sixteenth and seventeenth centuries, with an overall temperature range for the low-frequency component (>80-yr scales) of about 0.65 K. This amplitude can be compared with that in the smoothed and averaged low-resolution series in Fig. 2a, where the data used had been calibrated to temperatures by their original investigators. The size of the overall amplitudes of the series in Fig. 2a (0.6–0.9 K) and Fig. 2b (0.65 K, blue curve) indicates that the reconstructed multicentennial variability in Fig. 2b has not been inflated by the standardization, the wavelet processing and the subsequent calibration technique. The peaks in medieval times are at the same level as much of the twentieth century, although the post-1990 warmth seen in the

instrumental data (green curve in Fig. 2b) appears to be unprecedented.

Changes in radiative forcing due to variability in solar irradiance, the amount of aerosols from volcanic eruptions and greenhouse gas concentrations have been important agents causing climatic variability in the past millennium^{1,8,20–25}. Reconstructions of the temporal evolution of these variables have been used to drive climate models, ranging from simple energy balance models to fully coupled atmosphere–ocean general circulation models^{8,13,14,20,22,23,25}. The Northern Hemisphere temperature series obtained from such an experiment with the coupled model ECHO-G^{13,14,25,26} for the AD 1000–1990 period (Fig. 2c) is qualitatively remarkably similar to our multi-proxy reconstruction, although the model shows a somewhat larger variability. Particularly strong qualitative similarity is seen at timescales longer than centennial (compare the two wavelet transform patterns, Fig. 2e, f), but some high-frequency details are also in good agreement (for example, the double peaks around AD 1000 and 1100, and two short cooling periods at AD 1460 and 1820). The model's variability before the twentieth century is largely determined by the combined effects of the reconstructed solar and volcanic forcing (that is, natural forcing) that was used in the integration^{13,25}, whereas the notably strong warming in the model data after AD 1900 is largely due to rapidly increasing concentrations of anthropogenic greenhouse gases^{13,25,27}. The similarity between our reconstruction and the ECHO-G model series supports the case of a rather pronounced hemispheric low-frequency temperature variability resulting from the climate system's response to natural

Table 1 Summary information for low-resolution proxy series

No.*	Site†	Proxy type and interpretation‡	Dating type (max. dating error)	First, last year AD#	Sample resolution (no. of samples)§	Variance fraction (%) in >80-yr scales	
						Proxy	NH¶
1	Agassiz Ice Cap, Ellesmere Island, Canada, 81° N, 73° W	Ice melt, layer stratigraphy, summer <i>T</i>	Annual layer, theoretical model, ash layers (±10%)	–9, 1961	5 (393)	2.4	0.3
2	GRIP borehole, Greenland, 73° N, 38° W	Borehole <i>T</i> , annual <i>T</i>	Theoretical model	–300, 1996	50 (40)	89.2	12.5
3	Conroy Lake, Eastern US, 46° N, 68° W	Pollen, lake sediments, summer <i>T</i>	Annual laminations	35, 1968	7–40 (68)	69.2	9.7
4	Chesapeake Bay, Eastern US, 38° N, 76° W	Shells, Mg/Ca variation, spring <i>T</i>	¹⁴ C (±160 yr), ²¹⁰ Pb and ¹³⁷ Cs	5, 1995	1–65 (427)	30.8	4.3
5	N. Sargasso Sea, 33° N, 57° W	Foraminifera, δ ¹⁸ O, annual SST	¹⁴ C (±160 yr)	–300, 1925	50–150 (30)	65.0	9.1
6	NE Caribbean Sea, 18° N, 67° W	Foraminifera, δ ¹⁸ O, annual SST	¹⁴ C (±200 yr)	133, 1950	30–160 (26)	99.6	14.0
7	Lake Tsolmajavri, N. Finland, 68° N, 22° E	Diatoms, lake sediments, summer <i>T</i>	¹⁴ C (±169 yr)	–300, 1950	~70 (29)	61.5	9.9
8	Soylegrotta Cave, Mo i Rana, N. Norway, 66° N, 13° E	Stalagmite, δ ¹⁸ O, annual <i>T</i>	²³⁰ Th/ ²³⁴ U (±46 yr)	–300, 1938	~20 (103)	47.3	7.6
9	Shihua Cave, Beijing, China, 40° N, 116° E	Stalagmite, layer thickness, summer <i>T</i>	Layer counting (±5 yr)	–300, 1985	1 (1985)	32.2	5.2
10	China, 20–42° N, 80–130° E	Composite of 9 records, annual <i>T</i>	Layer counting, ¹⁴ C (±200 yr)	0, 1990	10 (200)	89.8	14.5
11	Arabian Sea, 18° N, 58° E	% <i>Globigerina bulloides</i> , up-welling, monsoons, summer and winter <i>T</i>	¹⁴ C (±100 yr)	–300, 1986	15–180 (63)	78.6	12.7

*Numbered as in Fig. 1.

†References and further site details are given in Supplementary Information.

‡*T*, temperature; SST, sea surface temperature.

#Negative numbers mean years before AD 1.

§Sample resolution in years. Number of samples after year AD 1 given in parentheses.

||For each proxy, the fraction of its total variance that occurs in its >80-yr scales (wavelet scales 10–22, see Methods) is listed.

¶For each proxy, its contribution (in %) to the total variance in the low-frequency component (wavelet scales 10–22) of the Northern Hemisphere (NH) temperature reconstruction is listed. These contributions are determined by the values in the second to last column. The resulting contributions by data from different regions are geographically well distributed; Arctic 12.9, Eastern US 14.0, Fennoscandia 17.6, China 19.7, Oceans 35.8. Furthermore, the five proxies that are annual mean temperature indicators (nos 2, 5, 6, 8, 10) together contribute 57.8% of the variance in the low-frequency component of the NH reconstruction.

changes in radiative forcing.

As in all palaeoclimate reconstructions, there are uncertainties in this one. Efforts to quantify uncertainties in the low-frequency component are illustrated with blue bands of different colours in Fig. 2d. Accounting for the total uncertainty, the cooling from the maxima around AD 1000 and 1100 to the minimum near AD 1600 is between 0.3 K and 1.1 K for the >80-yr component. Only the smallest value is similar to the corresponding temperature change in previous multi-proxy reconstructions^{1–4,7}. (Compare with the reconstruction of Mann *et al.*^{1,2}, which is also shown in Fig. 2d (orange curve, yellow uncertainty range) after smoothing to show >80-yr variability.) In contrast, the warming trend depicted by our reconstruction in the past 500 yr is entirely consistent with estimates of Northern Hemisphere ground surface temperature (GST) trends from temperature measurements in nearly 700 boreholes¹². Although trends at individual borehole sites are very noisy, hemispheric mean GST trends are robust to different aggregation and weighting schemes¹² (illustrated by the two brown curves in Fig. 2d), and show a warming of ~1 K during the interval AD 1500–2000, of which about half fell in the twentieth century.

There are several reasons for the notable differences between our and previous multi-proxy reconstructions. The reconstruction of

Mann and Jones⁷ has a large amount of data in common with ours, but these workers combined tree-ring data with decadal resolved proxies without any separate treatment at different timescales. Furthermore, our data set contains some centennially resolved data from the oceans, while Mann and Jones used only annually to decadal resolved data from continents or from locations very near the coast. Different calibration methods (regression in the work of Mann and Jones versus variance scaling in this study) are another reason. Possible explanations for the small low-frequency variability in the multi-proxy reconstruction of Mann *et al.*^{1,2} have recently been discussed in detail elsewhere¹⁴.

Our multi-proxy reconstruction achieves a relatively large multi-centennial variability by separately considering the information in tree-ring data and low-resolution proxy records. Much more effort, however, is needed to develop new proxy data series and to determine how proxy data should best be chosen, combined and calibrated. The wavelet approach allows for separate weighting of proxies at different timescales. A goal for further research could be to determine how such weighting should be undertaken, while simultaneously taking spatial representativity into account. New experiments with distorted ‘pseudoproxy’ data^{14,28,29} from forced general circulation model runs, using various types of distortions,

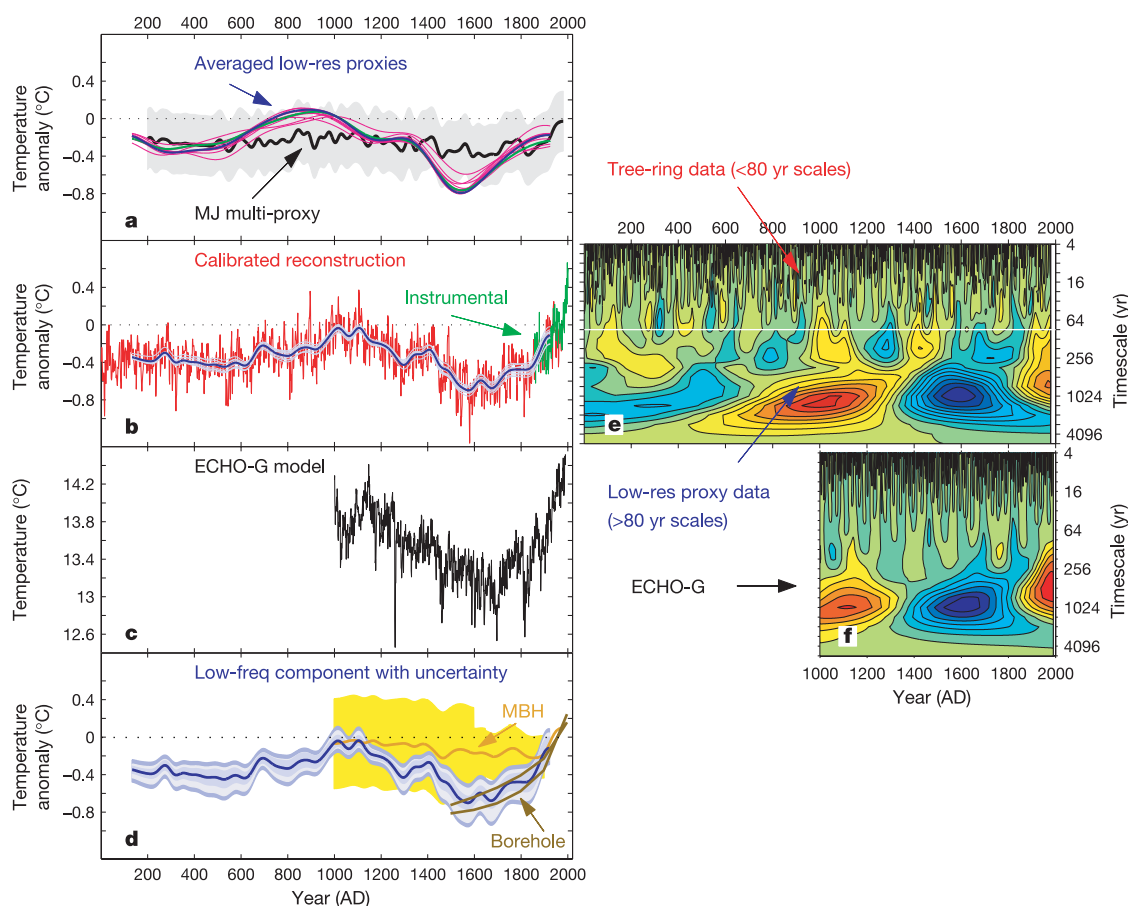


Figure 2 Estimations of Northern Hemisphere mean temperature variations. **a**, Previous multi-proxy reconstruction (MJ: from Mann and Jones⁷, black line) with ± 2 s.d. uncertainty (grey shading), and various averages (blue, green, magenta; see Methods) of smoothed low-resolution temperature indicators (nos 2–10 in Table 1). **b**, Our multi-proxy reconstruction AD 1–1979 (red) with its >80-yr component AD 133–1925 (blue) and its jack-knifed estimates (light blue), and the instrumental record¹⁹ (green). **c**, Forced ECHO-G model^{13,14,26}. **d**, Low-frequency component of multi-proxy reconstruction in **b** (blue curve) with confidence intervals for three separate uncertainties. The innermost medium-blue band shows the uncertainty due to variance among the low-resolution proxy

series. The two outer bands show the uncertainty in the variance scaling factor (light blue) and the constant adjustment (outermost blue band) separately. Each uncertainty is illustrated with an approximate 95% confidence interval (see Supplementary Information). Also shown are ground surface temperatures estimated from boreholes¹² with their uncertainty interval (brown; see Methods), and the >80-yr component of a previous multi-proxy reconstruction (MBH: from Mann *et al.*^{1,2}, orange) with ± 2 s.d. uncertainty (yellow). **e**, **f**, Wavelet transform (see Methods) of the series in **b** and **c** where warm/cold anomalies are red/blue. Temperature anomalies in **a**, **b** and **d** are with respect to the 1961–90 average.

should be useful in this context.

We find no evidence for any earlier periods in the last two millennia with warmer conditions than the post-1990 period—in agreement with previous similar studies^{1–4,7}. The main implication of our study, however, is that natural multicentennial climate variability may be larger than commonly thought, and that much of this variability could result from a response to natural changes in radiative forcings. This does not imply that the global warming in the last few decades^{15,19} has been caused by natural forcing factors alone, as model experiments that use natural-only forcings fail to reproduce this warming^{15,23,27,30}. Nevertheless, our findings underscore a need to improve scenarios for future climate change by also including forced natural variability—which could either amplify or attenuate anthropogenic climate change significantly. □

Methods

Wavelet transformation

The wavelet transform¹¹ (WT) decomposes a time series into time–frequency space, enabling the identification of both the dominant modes of variability and how those modes vary with time. We used the Mexican hat¹¹ wavelet, after linear interpolation to annual resolution, to decompose each proxy series into 22 wavelet ‘voices’. These are 22 separate time series, where each depict the variability in the original series at each of the timescales near the Fourier periods 4, $4\sqrt{2}$, 8, $8\sqrt{2}$, ..., $4,096\sqrt{2}$ yr. This ranges from the smallest possible scale for the Mexican hat wavelet up to a few times the length of the 2,000-yr-long time interval of interest. Before calculating the WT, padding with surrogate data to extend the proxy series from their last value to the year AD 2300 was applied to limit boundary effects at the longest timescales. As padding data we used the mean value for the last 50 yr with data in each series. For those records that do not have data back at least to 300 BC, similar padding was also made at the beginning of records.

Averaging of calibrated low-resolution temperature indicators

Nine low-resolution proxy series (Table 1) are available as calibrated local/regional temperature indicators (numbers 2–10). Five are annual mean (nos 2, 5, 6, 8, 10), one is a spring (no. 4) and three are summer (nos 3, 7, 9) temperature indicators. Multicentennial variability (highlighting >340-yr scales) in these series was extracted by first applying the WT to each series and then the inverse WT only for wavelet voices 14–22. Timescales of >340 yr are long enough to ensure that only information at frequencies lower than the Nyquist frequency for the most poorly resolved proxy series are used. Figure 2a shows arithmetic (green) and cos-latitude weighted (blue) averages of all nine indicators. Jack-knifed estimates of the cos-latitude weighted averages, where each of the nine indicators were removed one at a time, are also shown (magenta). These series are plotted in Fig. 2a for the period AD 133–1925 (when all have data) such that their mean values agree with the Mann and Jones reconstruction⁷ in the interval AD 200–1925. Minor differences between the arithmetic and cos-latitude weighted series indicate robustness to weighting choice. The similarity among the jack-knifed estimates indicates that the overall temperature range is relatively little affected by the removal of any individual indicator.

Timescale-dependent reconstruction

The multi-proxy reconstruction in Fig. 2b was constructed as follows. First, the data set of 19 proxy series in Fig. 1 was divided into two half-hemispheric subsets (western and eastern half). After wavelet transformation of each individual proxy series into 22 voices as described above (here using standardized series obtained by subtracting the mean and dividing by the standard deviation, after linear interpolation to annual resolution), a scale-by-scale averaging of wavelet voices was undertaken. For each wavelet scale from 1 to 9 (Fourier timescales <80 yr), the voices from the individual tree-ring series were averaged and, similarly, for each wavelet scale from 10 to 22 (Fourier timescales >80 yr) the corresponding voices from the individual low-resolution proxy series were averaged. Merging of the tree-ring-data average voices 1–9 with the low-resolution-data average voices 10–22 gave one complete WT with voices 1–22 for each half-hemispheric subset. The average of these two WTs then gave the whole-hemispheric WT shown in Fig. 2e. The rationale for dividing into half-hemispheric subsets is similar to that used in gridding: namely, to ensure that regions with many data series are weighted the same as regions (of the same size) with few data series. However, for our data set, little difference would be obtained if all data were placed in one whole-hemispheric set. A dimensionless Northern Hemisphere temperature reconstruction was obtained by calculating the inverse transform. Finally, this time series was calibrated by adjusting its variance and mean value to be the same as in the instrumental data¹⁹ in the overlapping period AD 1856–1979 (Fig. 2b). Separate estimates of uncertainties due to variance among individual low-resolution proxy data and in the determination of the variance scaling factor and constant adjustment term are explained in Supplementary Information.

Plotting GST trends

The two curves for GST trends¹² shown in Fig. 2d correspond to the upper and lower bounds for different area- and occupancy-weighted GST reconstructions, as illustrated in

figure 3 of ref. 12. To help visual comparison with our multi-proxy reconstruction, the GST trends are plotted with reference to the AD 1961–90 average in Fig. 2d. This was achieved by shifting the GST trend-curves such that they cross the zero level at 1958.5, which is the time when the twentieth century linear trend of the Northern Hemisphere instrumental surface air temperature¹⁹ (anomalies with respect to AD 1961–90) crosses the zero level. (The same technique was used in figure 5 of ref. 12.)

Received 21 July; accepted 9 December 2004; doi:10.1038/nature03265.

- Mann, M. E., Bradley, R. S. & Hughes, M. K. Global-scale temperature patterns and climate forcing over the past six centuries. *Nature* **392**, 779–787 (1998).
- Mann, M. E., Bradley, R. S. & Hughes, M. K. Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophys. Res. Lett.* **26**, 759–762 (1999).
- Jones, P. D., Briffa, K. R., Barnett, T. P. & Tett, S. F. B. High-resolution palaeoclimatic records for the last millennium: interpretation, integration and comparison with General Circulation Model control-run temperatures. *Holocene* **8**, 455–471 (1998).
- Crowley, T. J. & Lowery, T. S. How warm was the Medieval warm period? *Ambio* **29**, 51–54 (2000).
- Briffa, K. R. Annual climate variability in the Holocene: interpreting the message of ancient trees. *Quat. Sci. Rev.* **19**, 87–105 (2000).
- Esper, J., Cook, E. R. & Schweingruber, F. H. Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability. *Science* **295**, 2250–2253 (2002).
- Mann, M. E. & Jones, P. D. Global surface temperatures over the past two millennia. *Geophys. Res. Lett.* **30**, 1820, doi:10.1029/2003GL017814 (2003).
- Jones, P. D. & Mann, M. E. Climate over past millennia. *Rev. Geophys.* **42**, doi:10.1029/2003RG000143 (2004).
- Bradley, R. S. *Paleoclimatology: Reconstructing Climates of the Quaternary* (Academic, San Diego, 1999).
- Esper, J., Frank, D. C. & Wilson, R. J. S. Climate reconstructions: Low-frequency ambition and high-frequency ratification. *Eos* **85**, 113 (2004).
- Mallat, S. *A Wavelet Tour of Signal Processing* (Academic, San Diego, 1999).
- Pollack, H. N. & Smerdon, J. E. Borehole climate reconstructions: Spatial structure and hemispheric averages. *J. Geophys. Res.* **109**, doi:10.1029/2003JD004163 (2004).
- González-Rouco, F., von Storch, H. & Zorita, E. Deep soil temperature as a proxy for surface air-temperature in a coupled model simulation of the last thousand years. *Geophys. Res. Lett.* **30**, 2116, doi:10.1029/2003GL018264 (2003).
- von Storch, H. *et al.* Reconstructing past climate from noisy proxy data. *Science* **306**, 679–682 (2004).
- Intergovernmental Panel on Climate Change. *Climate Change 2001: The Scientific Basis* (IPCC, Geneva, 2001).
- Briffa, K. R. *et al.* Low-frequency temperature variations from a northern tree ring density network. *J. Geophys. Res.* **106**, 2929–2941 (2001).
- Cook, E. R., Briffa, K. R., Meko, D. M., Graybill, D. A. & Funkhouser, G. The ‘segment length curse’ in long tree-ring chronology development for palaeoclimatic studies. *Holocene* **5**, 229–237 (1995).
- Jones, P. D., Osborn, T. J. & Briffa, K. R. Estimating sampling errors in large-scale temperature averages. *J. Clim.* **10**, 2548–2568 (1997).
- Jones, P. D. & Moberg, A. Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001. *J. Clim.* **16**, 206–223 (2003).
- Crowley, T. J. Causes of climate change over the past 1000 years. *Science* **289**, 270–277 (2000).
- Shindell, D. T., Schmidt, G. A., Mann, M. E., Rind, D. & Waple, A. Solar forcing of regional climate change during the Maunder Minimum. *Science* **294**, 2149–2152 (2001).
- Bauer, E., Claussen, M., Brovkin, V. & Huenerbein, A. Assessing climate forcings of the Earth system for the past millennium. *Geophys. Res. Lett.* **30**, 1276, doi:10.1029/2002GL016639 (2003).
- Bertrand, C., Loutre, M.-F., Crucifix, M. & Berger, A. Climate of the last millennium: a sensitivity study. *Tellus A* **54**, 221–244 (2002).
- Shindell, D. T., Schmidt, G. A., Miller, R. L. & Mann, M. E. Volcanic and solar forcing of climate change during the Maunder Minimum. *J. Clim.* **16**, 4094–4107 (2003).
- Zorita, E. *et al.* Climate evolution in the last five centuries simulated by an atmosphere-ocean model: global temperatures, the North Atlantic Oscillation and the Late Maunder Minimum. *Meteorol. Z.* **13**, 271–289 (2004).
- Legutke, S. & Voss, R. *The Hamburg Atmosphere-Ocean Coupled Circulation Model ECHO-G* (Technical Report 18, Deutsches Klimarechenzentrum, Hamburg, 1999).
- Widmann, M. & Tett, S. F. B. Simulating the climate in the last millennium. *PAGES News* **11**(2&3), 21–23 (2003).
- Mann, M. E. & Rutherford, S. Climate reconstruction using ‘Pseudoproxies’. *Geophys. Res. Lett.* **29**, 1501, doi:10.1029/2001GL014554 (2002).
- Zorita, E., González-Rouco, F. & Legutke, S. Testing the Mann *et al.* 1998 approach to paleoclimate reconstructions in the context of a 1000-yr control simulation with the ECHO-G coupled climate model. *J. Clim.* **16**, 1378–1390 (2003).
- Stott, P. A. *et al.* External control of 20th century temperature by natural and anthropogenic forcings. *Science* **290**, 2133–2137 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank H. von Storch, E. Zorita and F. González-Rouco for the ECHO-G data, and H. Pollack and J. Smerdon for borehole data. All these persons and J. Esper, J. Luterbacher and M. Rummukainen are thanked for comments on early versions of the manuscript. We acknowledge financial support from the Royal Swedish Academy of Sciences, the Swedish Science Council and the Russian Foundation for Basic Research.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.M. (anders.moberg@misu.su.se).

Fractures as the main pathways of water flow in temperate glaciers

Andrew G. Fountain¹, Robert W. Jacobel³, Robert Schlichting² & Peter Jansson⁴

¹Departments of Geology and Geography, Portland State University, and

²Cleveland High School, Portland, Oregon 97202, USA

³Department of Physics, St Olaf College, Northfield, Minnesota 55057, USA

⁴Department of Physical Geography and Quaternary Geology, Stockholm University, SE-106 91, Stockholm, Sweden

Understanding the flow of water through the body of a glacier is important, because the spatial distribution of water and the rate of infiltration to the glacier bottom is one control on water storage and pressure, glacier sliding and surging, and the release of glacial outburst floods^{1–3}. According to the prevailing hypothesis, this water flow takes place in a network of tubular conduits^{4,5}. Here we analyse video images from 48 boreholes drilled into the small Swedish glacier Storglaciären, showing that the glacier's hydrological system is instead dominated by fractures that convey water at slow speeds. We detected hydraulically connected fractures at all depths, including near the glacier bottom. Our observations indicate that fractures provide the main pathways for surface water to reach deep within the glacier, whereas tubular conduits probably form only in special circumstances. A network of hydraulically linked fractures offers a simple explanation for the origin and evolution of the englacial water flow system and its seasonal regeneration. Such a fracture network also explains radar observations that reveal a complex pattern of echoes rather than a system of conduits. Our findings may be important in understanding the catastrophic collapse of ice shelves and rapid hydraulic connection between the surface and bed of an ice sheet.

The prevailing view of englacial water flow through tubular

conduits^{4,5} is supported by observations of caves on the margins of glaciers and explorations into moulins (a vertical shaft in a glacier)⁶. Whether such features are unique to the margin and to moulins is unclear. Observations in holes drilled to the glacier bed indicate the presence of englacial cavities^{7–9}, some of which appeared to be conduits. However, the hydraulic characteristics of these cavities were not defined and water movement, when detected, was slow¹⁰, much slower than expected theoretically.

To examine the physical and hydraulic characteristics of englacial water pathways we drilled into Storglaciären, a small (3 km²) polythermal glacier in Arctic Sweden¹¹. The surface cold layer typically does not extend below about 40 m, and the maximum ice depth is 250 m (ref. 12). Using high-pressure hot water we drilled holes, over three seasons (July–August), in the upper part of the ablation zone of the glacier (Fig. 1). Each hole was drilled until we intersected either an efficient hydraulic connection or the glacier bottom. This strategy avoided hydraulic interference resulting from multiple englacial connections within the same hole. Drilling with hot water through impermeable ice causes excess water to flow out of the hole. When the water level drops rapidly in the hole and remains below visual range (~3 m), we presumed that a hydraulically efficient connection was made. A submersible camera imaged the geometry and orientation of the structure at the point of drainage. Tracers (ink, dye, teflon spheres) were injected via a vinyl tube to track fluid motion. In some holes we installed pressure transducers to record changes in water pressure. We also surveyed the subsurface with ice-penetrating radar to detect the presence of englacial features remotely.

Thirty-eight (79%) of our 48 holes intercepted an efficient englacial hydraulic passage and 80 cavities were imaged. Almost half of the holes had multiple cavities of which the uppermost must have been hydraulically isolated or poorly connected because the water level in the holes did not drop when the drill passed through them. We could not discern the geometry of about half of the cavities because the surrounding ice was clear, unlike the bubbly ice elsewhere, and so the walls provided insufficient reflection from the camera lights. Of the remaining 44 cavities, 36 (80%) were fracture-

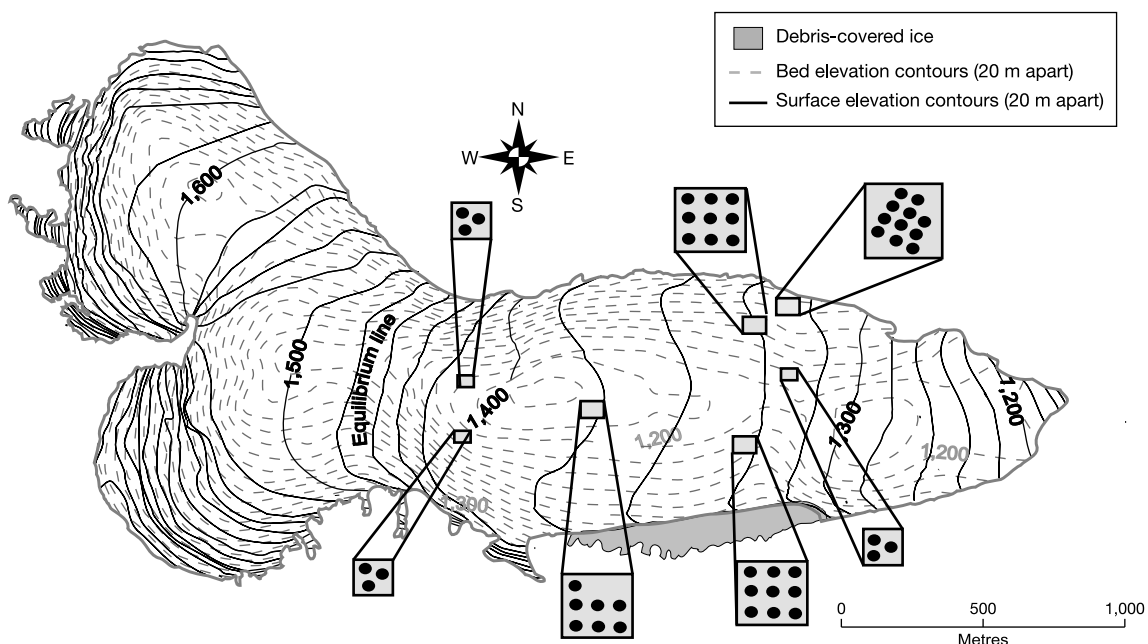


Figure 1 Map of Storglaciären showing the location of the drill sites. Four sites were composed of at least seven holes, each 10 m apart in a grid plan, and three sites were composed of three holes, each ~20 m apart in a triangular plan. All sites were drilled in

the over-deepened section of the glacier with the exception of the first set in the upper right.

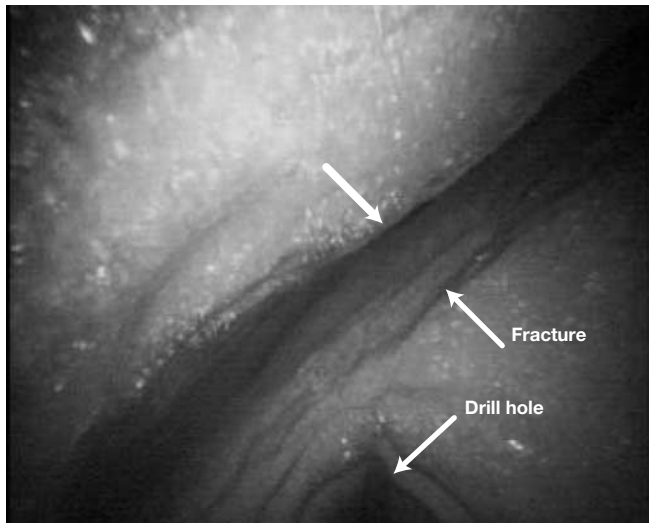


Figure 2 Video image of an englacial fracture. The fracture width is about 4 cm and the continuation of the vertical drill hole is identified. The camera is tilted obliquely downward towards the fracture.

like features, two (4%) were circular in cross-section and the remaining six (16%) were amorphous and appeared to be the intersections of multiple fractures.

The fracture-like features had near-vertical ($\sim 70^\circ$) subparallel walls (Fig. 2). Also, we made multiple drillings into the same feature over a horizontal distance of several metres and found a similar steep dip. All of these features appear identical to images taken in water-filled surface crevasses. Therefore, we consider them to be englacial fractures. Fractures were detected at depths from near the surface to 96% of the local glacier thickness; the deepest observed fracture was 131 m below the surface. Fracture orientation (strike) was generally parallel to strain-related features within the surface structure of Storglaciären including clear ice bands inferred to be refrozen crevasses^{7,8}. A few strikes, particularly those at depth, were

rotated tens of degrees and whether this represents a different structural control at depth or conjugate fractures is unclear. The width of fracture openings varied from 0.3 to 20 cm with a median of 4 cm and did not correlate with depth. These widths are maximum values because the openings must have been enlarged as the warm drill water drained into the fractures. Over distances of ~ 100 m the hydraulic gradient in the fracture system mirrors that of the surface slope of the glacier.

Water flow was observed in only ten fractures. Flow speeds were measured from particles moving past the camera view and from tracer tests between adjacent holes. They ranged from 0.5 to 4 cm s^{-1} , with typical speeds of $1\text{--}2 \text{ cm s}^{-1}$. These speeds are slow compared to expectations from models of flow in tubular conduits^{4,5}. Reynolds numbers indicate that the flow is laminar. Sometimes holes 10–20 m apart were hydraulically connected via the fractures, as indicated by tracers leaking from one hole to another or by pumping water into one hole and observing a water level response in another. In one case, the connection depth in a pair of holes differed by 91 m. A flow test between holes, where water was pumped into one hole creating a constant head difference between holes, showed that the measured flow speed was 1% of that predicted on the basis of laminar flow between parallel plates. For the majority of fractures, which showed no observable flow, either flow speeds were below detection, or the fractures were side passages to the main flow routes.

After drilling 47 holes and not intersecting any tubular conduits, we searched in a field of moulins, where conduits are known to occur. At a depth of 42 m we intersected a conduit with a diameter of 10 cm and a water flow speed of $\sim 10 \text{ cm s}^{-1}$. This feature was the only 'traditional' conduit found. It appeared to be gently sloping with an orientation parallel to the strike of nearby surface crevasses, which were perpendicular to the direction of ice flow.

To detect englacial cavities remotely we completed 247 radar (50 and 100 MHz) profiles covering ~ 25 km of surface travel. The profiles formed a series of closely spaced grids over the drill holes. As expected, the profiles showed a myriad of englacial echoes typical of radar surveys in temperate ice^{13–15}. Although elongated echoes are not common in temperate ice, they are observed occasionally¹⁶ and

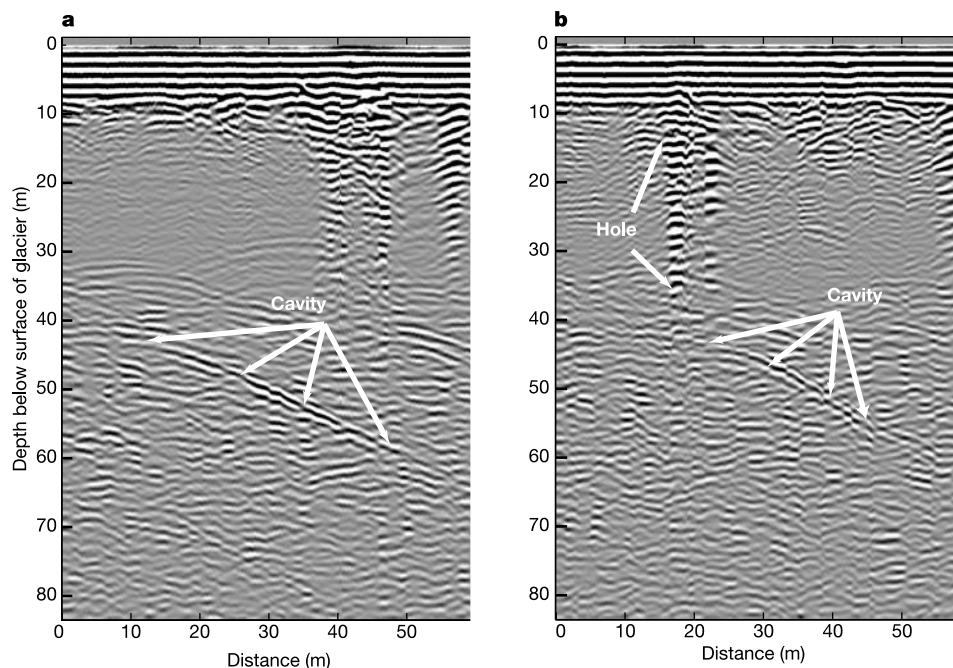


Figure 3 Radar (50-MHz) reflections from the glacier interior showing a dipping reflector before and after drilling. **a**, The radar reflections before drilling with a hot water drill; **b**, the reflections after intersecting the top of the reflector at a depth of 38 m near a distance of

18 m. The reflector, inferred to be a fracture, dips to the north (right), with the strongest reflection between depths of ~ 40 and ~ 55 m.

we found a few on Storglaciären. One prominent dipping reflector was explored (Fig. 3). It originated at ~40 m depth and was ~20 m long when projected onto the glacier surface. We drilled to the inferred feature and intersected a hydraulically connected fracture at 38 m depth (Fig. 3b). The hole did not intersect the brightest segment of the fracture but probably encountered the upper part, which is not well depicted by the radar. The camera imaged a fracture, ~4 cm wide, dipping in the same direction as indicated by the radar.

The drilling and radar results point to copious water-filled cavities within temperate ice. Images showed that 80% of these cavities were steeply dipping fractures, most of which were hydraulically connected, forming an englacial hydraulic system. Although a few englacial fractures have been observed previously^{7,10}, we show here that they are ubiquitous and form an integrated hydrological network. The general absence of tubular conduits is surprising. Although one could argue that the probability of intersecting conduits with a drill is small¹⁰ compared to the probability of intersecting a planar fracture, even a steeply dipping one, the ubiquitous presence of hydraulically connected fractures is undeniable. Our results support previous work that suggests an englacial hydraulic system that is not dominated by tubular conduits¹⁷.

The presence of a network of hydraulically connected fractures may resolve several important questions in glacier hydrology. First, it has been unclear how tubular conduits form and how surface water is routed into them. In our view, surface crevasses provide ready entry^{3,18} and a crevasse that is nearly filled with water can propagate to significant depths and perhaps to the bottom of the glacier^{19,20}. Second, it was not clear how conduits survive a winter of closure and deformation due to ice shear, eventually to regenerate the following spring with the onset of melt²¹. Fracturing resolves both issues because it is a continuous process that generates passageways through the ice. Likewise, a ubiquitous fracture network helps to explain the often puzzling behaviour of water levels observed in many holes drilled all the way down to the bed that do not clearly correspond to diurnal surface forcing or the basal response^{17,22}. It seems likely that these holes inadvertently established multiple hydraulic connections between englacial and subglacial systems, and the variations in water level are the result of interacting systems. Consequently, water level variations in such holes provide dubious information about subglacial water pressures^{22,23}. Finally, the myriad of echoes detected by radar surveys, rather than the bright and elongated reflector expected from a water-filled conduit¹⁶, probably originate as reflections from fractures that overlie echoes from other fractures.

Our results are generally applicable to other glaciers. Although most of our holes were drilled in the gently sloping over-deepened part of Storglaciären, our first set of holes was drilled along the glacier margin where the surface slopes steeply towards the valley wall. Observations of fractures in this first set were no different from those in holes in the over-deepened section and therefore no influence of glacier slope or depth can be detected. Also we reviewed video tapes from holes drilled during a previous project on Kennicott glacier, a temperate glacier in southern Alaska²⁴. These holes also terminated englacially, intersected fractures, and were hydraulically connected. We conclude that fracture networks are common on temperate and polythermal glaciers and are probably not confined to any one region within a glacier. We infer that englacial fractures will also be common in the accumulation zone of glaciers.

The origin of englacial fractures is unclear but advection of surface crevasses²⁵ and *in situ* generation are possible mechanisms. *In situ* generation was observed at Bench glacier (Alaska) when an englacial fracture appeared in a hole wall, at 137 m depth, during a 10-day interval between observations (J. Harper, personal communication). Hydraulic linkages between fractures may develop as

new fractures intersect relic, water-filled ones to form continuous water pathways. Because of the slow water speed, viscous shear heating and melting of fracture walls should not play a significant role in controlling water flow. Consequently, the hydraulics of englacial water flow in a fracture-dominated glacier will differ from expectations based on flow in conduits.

Tubular conduits probably develop in special circumstances where the water flux is sufficiently large to initiate shear heating and melting of the fracture walls. A surface stream discharging into a crevasse might create such conditions. Localized flow enhancement develops at the expense of flow elsewhere in the fracture, resulting in conduit formation. This process of destabilizing a fracture network to form a conduit may be analogous to that for destabilizing sheet flow at the base of a glacier in favour of conduits²⁶. Other fractures will drain to conduits where they intersect.

The hypothesis of fracture networks in glaciers may be important in understanding other glaciological phenomena such as the catastrophic collapse of ice shelves²⁷ and rapid hydraulic connection (and dynamic response) between the surface and bed of an ice sheet²⁸. □

Received 11 May; accepted 13 December 2004; doi:10.1038/nature03296.

1. Iken, A. & Truffer, M. The relationship between subglacial water pressure and velocity of Findelengletscher, Switzerland, during its advance and retreat. *J. Glaciol.* **43**, 328–338 (1997).
2. Paterson, W. S. B. *The Physics of Glaciers* 103–172 (Pergamon, Oxford, 1996).
3. Fountain, A. G. & Walder, J. S. Water flow through temperate glaciers. *Rev. Geophys.* **36**, 299–328 (1998).
4. Shreve, R. L. Movement of water in glaciers. *J. Glaciol.* **11**, 205–214 (1972).
5. Röthlisberger, H. Water pressure in intra- and subglacial channels. *J. Glaciol.* **11**, 177–203 (1972).
6. Holmlund, P. Internal geometry and evolution of moulins, Storglaciären, Sweden. *J. Glaciol.* **34**, 242–248 (1988).
7. Pohjola, V. A. TV-video observations of englacial voids in Storglaciären. *J. Glaciol.* **40**, 231–240 (1994).
8. Harper, J. T. & Humphrey, N. F. Borehole video analysis of a temperate glacier's englacial and subglacial structure: Implications for glacier flow models. *Geology* **23**, 901–904 (1995).
9. Copland, L., Harbor, J. & Sharp, M. Borehole video observation of englacial and basal ice conditions in a temperate valley glacier. *Ann. Glaciol.* **24**, 277–282 (1997).
10. Hooke, R. L. & Pohjola, V. A. Hydrology of a segment of a glacier situated in an overdeepening, Storglaciären, Sweden. *J. Glaciol.* **40**, 140–148 (1994).
11. Jansson, P. Dynamics and hydrology of a small polythermal valley glacier. *Geograf. Ann. Ser. A* **78**, 171–180 (1996).
12. Pettersson, R., Jansson, P. & Holmlund, P. Cold surface layer thinning on Storglaciären, Sweden, observed by repeated ground penetrating radar surveys. *J. Geophys. Res.* **108**, doi:10.1029/2003JF000024 (2003).
13. Fountain, A. G. & Jacobel, R. W. Advances in ice radar studies of a temperate alpine glacier, South Cascade Glacier, Washington, U.S.A. *Ann. Glaciol.* **24**, 303–308 (1997).
14. Arcone, S. A., Lawson, D. E. & Delaney, A. J. Short-pulse radar wavelet recovery and resolution of dielectric contrasts within englacial and basal ice of Matanuska Glacier, Alaska, U.S.A. *J. Glaciol.* **41**, 68–86 (1995).
15. Murray, T., Stuart, G. W., Fry, M., Gamble, N. H. & Crabtree, M. D. Englacial water distribution in a temperate glacier from surface and borehole radar velocity analysis. *J. Glaciol.* **46**, 389–398 (2000).
16. Stuart, G., Murray, T., Gamble, N., Hayes, K. & Hodson, A. Characterization of englacial channels by ground-penetrating radar: An example from austre Broggerbreen, Svalbard. *J. Geophys. Res. Solid Earth* **108**, doi:10.1029/2003JB002435 (2003).
17. Hock, R., Iken, A. & Wängler, A. Tracer experiments and borehole observations in the over-deepening of Aletschgletscher, Switzerland. *Ann. Glaciol.* **28**, 253–260 (1999).
18. Stenborg, T. Some viewpoints on the internal drainage of glaciers. *Int. Ass. Hydrol. Sci.* **95**, 117–129 (1973).
19. de Robin, G. Q. Depth of water-filled crevasses that are closely spaced. *J. Glaciol.* **13**, 543 (1974).
20. Van der Veen, C. J. Fracture mechanics approach to penetration of surface crevasses on glaciers. *Cold Regions Sci. Technol.* **27**, 31–47 (1998).
21. Gordon, S. et al. Seasonal reorganization of subglacial drainage inferred from measurements in boreholes. *Hydrol. Process.* **12**, 105–133 (1998).
22. Gordon, S. et al. Borehole drainage and its implications for the investigation of glacier hydrology: experiences from Haut Glacier d'Arolla, Switzerland. *Hydrol. Process.* **15**, 797–813 (2001).
23. Humphrey, N., Kamb, B., Fahnestock, M. & Engelhardt, H. Characteristics of the bed of the lower Columbia Glacier, Alaska. *J. Geophys. Res.* **98**, 837–846 (1993).
24. Anderson, S. P. et al. Integrated hydrologic and hydrochemical observations of Hidden Creek Lake jökullsluip, Kennicott Glacier, Alaska. *J. Geophys. Res.* **108**, doi:10.1029/2002JF000004 (2003).
25. Pohjola, V. A. *Ice Dynamical Studies of Storglaciären, Sweden* PhD thesis, Uppsala Univ. (1993).
26. Walder, J. S. Stability of sheet flow of water beneath temperate glaciers and implications for glacier surging. *J. Glaciol.* **28**, 273–293 (1982).
27. Scambos, T. A., Hulbe, C., Fahnestock, M. & Bohlander, J. The link between climate warming and break-up of ice shelves in the Antarctic Peninsula. *J. Glaciol.* **46**, 516–530 (2000).
28. Zwally, H. J. et al. Surface melt-induced acceleration of Greenland Ice-Sheet flow. *Science* **297**, 218–222 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Frödin-Nyman, M. Nyman, R. Pettersson, and R. Hock. J. Walder and R. LeB. Hooke provided suggestions during the preparation of this manuscript. This work was supported by the Arctic Natural Sciences Section within the Office of Polar Programs of the US National Science Foundation. P.J.'s participation was funded by the Swedish Research Council. The Stockholm University, Tarfala Research Station provided excellent facilities and field assistance.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.G.F. (andrew@pdx.edu).

Abrupt rise in atmospheric CO₂ overestimates community response in a model plant–soil system

John N. Klironomos¹, Michael F. Allen², Matthias C. Rillig³, Jeff Piotrowski³, Shokouh Makvandi-Nejad¹, Benjamin E. Wolfe¹ & Jeff R. Powell¹

¹Department of Integrative Biology, University of Guelph, Guelph, Ontario N1G 2W1, Canada

²Center for Conservation Biology, University of California, Riverside, California 92521, USA

³Division of Biological Sciences, The University of Montana, Missoula, Montana 59812, USA

Attempts to understand the ecological effect of increasing atmospheric CO₂ concentration, [CO₂], usually involve exposing today's ecosystems to expected future [CO₂] levels^{1,2}. However, a major assumption of these approaches has not been tested—that exposing ecosystems to a single-step increase in [CO₂] will yield similar responses to those of a gradual increase over several decades³. We tested this assumption on a mycorrhizal fungal community over a period of six years. [CO₂] was either increased abruptly, as is typical of most [CO₂] experiments, or more gradually over 21 generations. The two approaches resulted in different structural and functional community responses to increased [CO₂]. Some fungi were sensitive to the carbon pulse of the abrupt [CO₂] treatment. This resulted in an immediate decline in fungal species richness and a significant change in mycorrhizal functioning. The magnitude of changes in fungal diversity and functioning in response to gradually increasing [CO₂] was smaller, and not significantly different to those with ambient [CO₂]. Our results suggest that studies may overestimate some community responses to increasing [CO₂] because biota may be sensitive to ecosystem changes that occur as a result of abrupt increases.

A major goal in climate change research is to predict the structure and functioning of ecological systems at a future date when the climate will be significantly different from that of today⁴. For example, atmospheric [CO₂] is expected to continue rising during the next century to concentrations of 550 p.p.m. or more⁴. A major research effort is under way to understand the changes that will occur in population, community and ecosystem structure and function in response to this increase in [CO₂]⁵.

A typical experimental approach used to predict the ecological effect is to expose current ecosystems to ambient and elevated atmospheric [CO₂] and compare the responses^{1,2,5}. A major assumption in such [CO₂] experiments is that exposing today's ecosystems to an abrupt increase in [CO₂] will yield structural and functional responses similar to those that would be observed by exposing the same ecosystems to a gradual increase of the same

magnitude over several decades. There is evidence that plants undergo microevolutionary changes in response to elevated [CO₂]^{6,7}, and that there are compositional changes in communities^{5,6}. Many organisms, especially those that grow quickly, will have gone through many generations, each one exposed to an incrementally higher [CO₂]. Subjecting organisms to a step increase in [CO₂] may exert a selective pressure on biota very different from the selective pressure exerted when each generation is subjected to an incrementally increased [CO₂]. Theory indicates that gradual increases in [CO₂] over many generations may elicit responses that are different from those following abrupt increases³. However, experimental support for this claim has been lacking.

Using a model experimental system, we tested the hypothesis that biodiversity and function would respond differently depending on whether a community is exposed to an abrupt or a gradual increase in [CO₂]. The arbuscular mycorrhizal symbiosis is ideal for testing this hypothesis for two main reasons. First, arbuscular mycorrhizal fungi (AMF) are affected indirectly by elevated atmospheric [CO₂], responding to changes in plant physiology and growth^{8–10}. These fungi are obligate biotrophs that are intimately associated with plant roots and depend directly on plant photosynthate as a source of carbon^{11,12}. Increasing atmospheric [CO₂] often results in increased allocation of carbon to roots. This increased carbon availability can influence microbial interactions in the rhizosphere and the structure of the AMF community^{8,13–15}. Second, AMF and many of their plant hosts grow and reproduce quickly. This allows us to study the responses of several generations over a reasonably short period of time.

The experiment was conducted using *Bromus inermis* (a perennial C3 grass) and its associated mycorrhizal community from the Long-Term Mycorrhiza Research Site in Guelph, Ontario, Canada¹⁶. *B. inermis* was chosen because it is a common plant at the site, shows significant growth responses when associated with AMF¹⁶, and preliminary studies have demonstrated that the plant significantly increases its dependency on AMF for growth in an elevated atmospheric [CO₂] environment (data not shown). Plants and fungi were grown for 21 generations, each generation lasting 15 weeks (about 6 yr in total). Plants were subjected to one of three treatments: (1) 350 p.p.m. [CO₂] at each generation (here referred to as 'ambient'); (2) 550 p.p.m. [CO₂] at each generation (here referred to as 'abrupt'); or (3) 350 p.p.m. [CO₂] at the first generation, increasing by 10 p.p.m. at each subsequent generation, with the final generation exposed to 550 p.p.m. (here referred to as 'gradual') (see Methods for more details).

After each generation we determined the biodiversity and functioning of the resulting fungal community. For biodiversity, we identified AMF taxa following a trap-culture bioassay and determined AMF species richness. For functioning, we determined the ability of the AMF community to influence plant biomass. On the first and final generations, we also measured a number of other variables that are highly influenced by AMF, including the concentration of phosphorus in plant tissue, [P], root length, per cent mycorrhizal colonization, mycelium production, and soil aggregate size distribution and aggregate water stability (both measures of soil aggregation).

We did not find any evidence of plant genotype selection over the course of the experiment. We grew seedlings from all generations in a common environment (on a field-collected soil under greenhouse conditions) and did not observe significant differences in gross photosynthesis rates, above- or below-ground plant biomass, or their mycorrhization (data not shown). Furthermore, measures of plant photosynthesis under 550 p.p.m. [CO₂] did not differ between gradual and abrupt treatments (data not shown). However, we did observe significantly higher below-ground plant production in the abrupt treatment, which suggests that the plants were responding to changes in the soil environment. The plants were most probably responding to changes in the composition of the mycorrhizal

Table 1 Influence of increased [CO₂] on mycorrhizal structure and function

Response variables	Treatment		
	Ambient	Abrupt	Gradual
Generation 1			
Total plant biomass (g)	7.1 (1.8) ^a	11.6 (2.5) ^a	8.2 (1.3) ^a
Leaf [P]	30.3 (2.9) ^a	17.1 (1.8) ^b	29.9 (2.0) ^a
Root length (m)	297 (42) ^a	329 (45) ^a	290 (36) ^a
AMF root colonization (%)	42.7 (4.7) ^a	38.8 (3.3) ^a	37.8 (3.8) ^a
Hyphal length (m g ⁻¹ soil)	5.4 (0.4) ^a	2.9 (0.2) ^b	4.1 (0.4) ^{ab}
<i>Glomus</i> hyphae (m g ⁻¹ soil)	1.7 (0.4) ^a	1.4 (0.3) ^a	1.3 (0.2) ^a
<i>Acaulospora</i> hyphae (m g ⁻¹ soil)	0.2 (0.1) ^a	0.3 (0.1) ^a	0.1 (0.1) ^a
<i>Gigaspora</i> hyphae (m g ⁻¹ soil)	0.8 (0.1) ^a	0.2 (0.05) ^b	0.7 (0.1) ^a
<i>Scutellospora</i> hyphae (m g ⁻¹ soil)	0.6 (0.2) ^a	0.0 (0.0) ^b	0.9 (0.2) ^a
MWD (mm)	1.53 (0.04) ^a	1.52 (0.02) ^a	1.52 (0.03) ^a
WSA (%)	92.6 (0.8) ^a	90.1 (1.3) ^a	89.7 (1.1) ^a
Generation 21			
Total plant biomass (g)	6.8 (1.2) ^a	10.7 (2.0) ^a	7.8 (1.6) ^a
Leaf [P]	24.1 (1.9) ^{ab}	16.7 (2.0) ^a	28.3 (2.1) ^b
Root length (m)	250 (38) ^a	394 (42) ^b	282 (35) ^a
AMF root colonization (%)	38.2 (2.5) ^a	36.3 (3.1) ^a	34.9 (2.9) ^a
Hyphal length (m g ⁻¹ soil)	4.9 (0.4) ^a	2.4 (0.3) ^b	4.9 (0.3) ^a
<i>Glomus</i> hyphae (m g ⁻¹ soil)	2.1 (0.4) ^a	1.5 (0.3) ^a	1.5 (0.4) ^a
<i>Acaulospora</i> hyphae (m g ⁻¹ soil)	0.1 (0.1) ^a	0.2 (0.1) ^a	0.2 (0.1) ^a
<i>Gigaspora</i> hyphae (m g ⁻¹ soil)	0.5 (0.1) ^a	0.1 (0.03) ^b	0.8 (0.1) ^a
<i>Scutellospora</i> hyphae (m g ⁻¹ soil)	1.0 (0.03) ^a	0.2 (0.03) ^b	0.9 (0.2) ^a
MWD (mm)	1.53 (0.02) ^a	1.75 (0.05) ^b	1.45 (0.02) ^a
WSA (%)	91.9 (1.8) ^a	92.0 (0.7) ^a	91.0 (1.6) ^a

Values represent the mean, followed by standard error in parentheses. Different superscript letters (a, b) represent significant differences ($P < 0.05$) after ANOVA and Tukey post-hoc tests.

community (see below). The most carbon-demanding fungal taxa were lost in the abrupt treatment, possibly releasing the plants from a heavy carbon drain.

Per cent root colonization by AMF did not change in response to any [CO₂] treatment (Table 1). However, we detected significant treatment effects on AMF species richness. Average AMF species richness across the entire ambient treatment was significantly higher than the average species richness in the abrupt treatment (Fig. 1). Average species richness across generations in the gradual treatment

did not significantly differ from that in the ambient, but did differ from the abrupt treatment. At the end of the experiment (final generation 21), AMF species richness in the ambient and gradual treatments remained similar, and contained twice as many AMF taxa compared with the abrupt treatment.

More AMF taxa were lost when [CO₂] was raised abruptly than when it was raised gradually. However, not all taxa were equally vulnerable. The number of species observed from the genus *Glomus* was similar under the three treatments, but the number of species in the other genera (mainly *Gigaspora* and *Scutellospora*) declined sharply. On average, *Glomus* spp. represented 62% of all taxa in the ambient treatment and 67% of taxa in the gradual treatment. However, species of *Glomus* made up 93% of taxa in the abrupt treatment. In this latter treatment, the length of living hyphae in the soil also declined significantly for species of *Gigaspora* and *Scutellospora*, but remained constant for species of *Glomus* (Table 1). The decline in fungal species diversity in the abrupt treatment was observed after generation 1, indicating that species of *Gigaspora* and *Scutellospora* are sensitive to the abrupt carbon pulse.

The change in AMF species composition under the different [CO₂] treatments also resulted in differences in functioning. In the abrupt treatment, the elimination of many *Gigaspora* and *Scutellospora* species, and the reduction in hyphal biomass through the loss of these taxa, resulted in an AMF community that was better able to stimulate plant biomass compared with the other two treatments (Fig. 2). Also, total P, determined by multiplying [P] with plant biomass, did not differ significantly among treatments, suggesting the P available to plants was diluted in larger plants. Such a dilution of P may not necessarily indicate higher P demand by the plant. However, it does indicate that the resulting AMF community was less capable of stimulating plant tissue [P]. Species of *Gigaspora* and *Scutellospora* are known to produce the most extensive extraradical mycelia^{17,18}. In the present study, the highest mycelial production in the soil was observed in the ambient and gradual treatments, in which *Gigaspora* and *Scutellospora* species had their

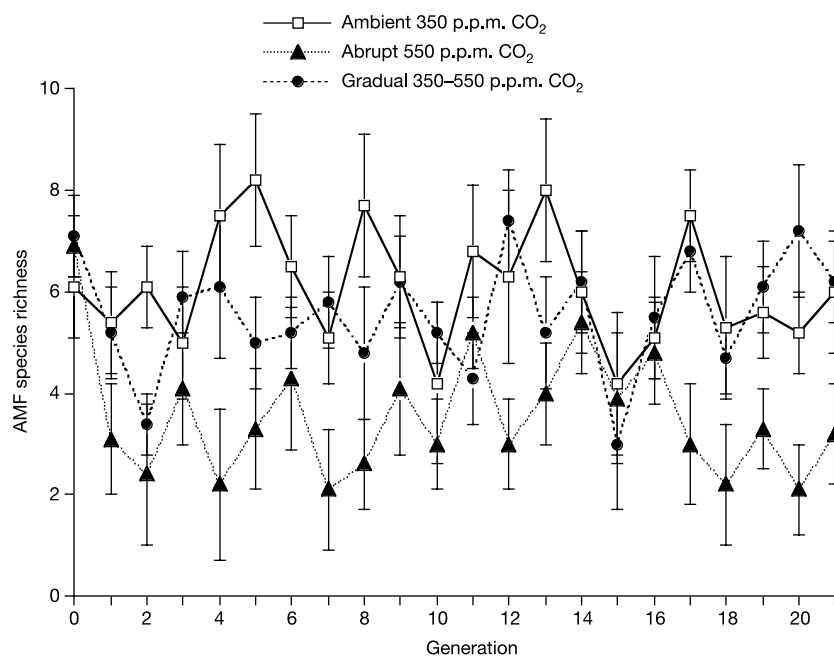


Figure 1 The effect of atmospheric [CO₂] on AMF species richness. Overall, mean richness was significantly higher (repeated-measures ANOVA (Tukey), $P = 0.001$) in the 'ambient' compared with the 'abrupt' treatment. Richness in the 'gradual' treatment did not significantly differ from the ambient treatment (repeated-measures ANOVA (Tukey), $P = 0.06$), but did differ from the abrupt treatment (repeated-measures ANOVA (Tukey),

$P = 0.001$). AMF species richness at generation 21 in the ambient and gradual treatments remained similar (ANOVA (Tukey), $P = 0.879$), and contained twice as many AMF taxa compared with the abrupt treatment (ANOVA (Tukey), $P = 0.001$). Points represent the mean and bars represent the standard error.

highest occurrence (Table 1). High investment in mycelial production may result in an improved ability to extract P from the soil and translocate it to the plant host, but it also results in a higher demand for carbon from the plant.

The different CO₂-mediated changes in AMF species composition had other functional consequences as well, such as in the formation of soil aggregates. AMF play a central role in this process¹⁹, which reduces soil losses associated with wind and water erosion and greatly influences other biological, physical and chemical soil processes. The soil of the *Glomus*-dominated AMF communities found in the abrupt treatment produced higher mean weight diameter (MWD) soil aggregates in conjunction with plant roots (Table 1), indicating a shift in soil aggregates towards greater mass in larger aggregate size classes (as had been found previously with step-increase experiments¹⁹). The absence of change in the percentage of water-stable aggregates (WSA, Table 1) indicated that this was not a transient response, but that aggregates had similar water stability irrespective of [CO₂] treatments. The soil aggregation results are most probably a response to the decrease in the relative abundance of *Gigaspora* and increased root lengths. *Gigaspora* has been shown to be the least effective AMF group at stimulating soil aggregation when associated with *B. inermis*²⁰.

The present results show that exposing plants to elevated [CO₂] could influence the structure and functioning of their associated AMF. This phenomenon has been shown in many other studies^{8–10,15} that have instantaneously raised the [CO₂] from ambient to ≥550 p.p.m. However, the present data also support the hypothesis that AMF responses to an abrupt increase in [CO₂] could differ from responses to a gradual increase if an initial loss of fungal taxa is not replaced. Global atmospheric [CO₂] is increasing at a rate (around 1.5 p.p.m. per year)⁴ that is much lower even than that of the gradual treatment. Under that treatment, AMF diversity and functioning were not significantly different from those under ambient conditions. As a result, we conclude that mycorrhizal functioning may not be as profoundly affected by increasing

atmospheric [CO₂] as is generally reported.

The structure and functioning of communities and ecosystems will alter in response to environmental change, such as increasing atmospheric [CO₂], warming and nitrogen enrichment. Determining the direction and magnitude of such responses is important because they are used to make decisions regarding current and future management of ecosystems⁴. Experimental efforts to determine the nature and magnitude of such responses typically use the abrupt step-increase approach. We cannot be certain that the differential responses to abrupt versus gradual increase in [CO₂] detected here, with this model system, will also be observed in other ecosystems. However, given that mycorrhizal fungi are ubiquitous^{11,12} and that their presence and composition in natural ecosystems can have profound effects on the structure and function of plant communities²¹, the structure of higher trophic levels^{22,23} and various ecosystem processes²⁴, there is little reason to think that such responses will be restricted to this model community. Also, the model system used here was fundamentally that of a plant–microbe trophic interaction. The observed responses were primarily a consequence of the sensitivity of certain AMF to a pulse of carbon at the start of the experiment. Will such a pulse also influence other interactions, such as plant–parasite and plant–insect interactions? At present, we cannot tell.

In conclusion, the abrupt approach resulted in a significant change in mycorrhizal diversity and functioning in the first generation, with little change after additional generations. It is not known whether the effects would be similar in an intact field experiment where fungal meta-community dynamics may come into play and mediate any local species extinctions. We therefore encourage research in other model systems and in intact ecosystems to quantify the proportion of observed responses that is a result of exposure to an abrupt increase in [CO₂]. Finally, our present experimental study arrives at evidence similar to the model predictions of ref. 3, in which plant carbon allocation would be overestimated using the abrupt approach. Both studies suggest that caution should be exercised in interpreting the magnitude of

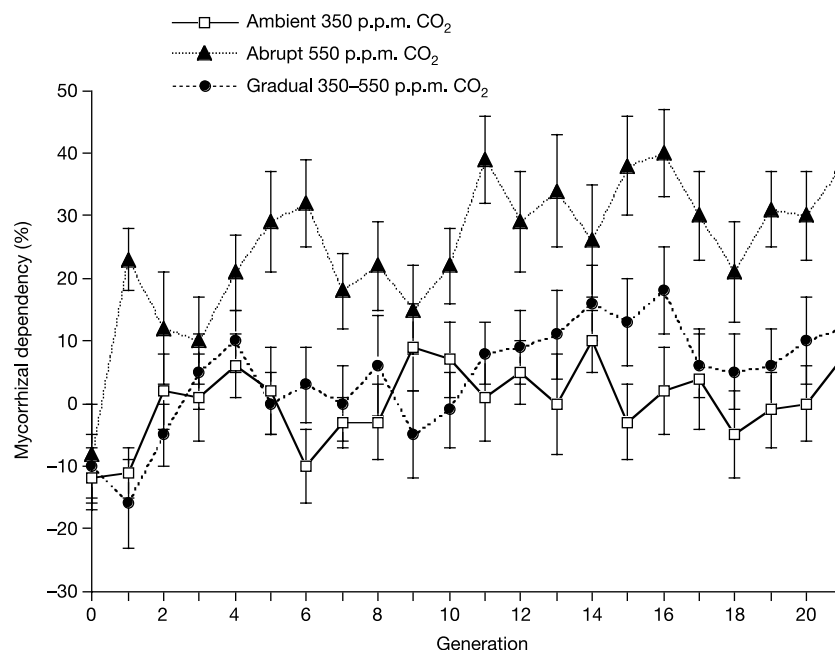


Figure 2 The effect of atmospheric [CO₂] on mycorrhizal dependency of *B. inermis*. Points represent the mean and bars represent the standard error. Mycorrhizal dependency was similar before the treatments (generation 0). Overall, mean mycorrhizal dependency was significantly higher (repeated-measures ANOVA (Tukey), $P = 0.001$) in the abrupt compared with the ambient and gradual treatments. Mean mycorrhizal dependency did

not differ between the latter two treatments (repeated-measures ANOVA (Tukey), $P = 0.71$). At generation 21, mycorrhizal dependency was significantly higher in the 'abrupt' treatment (ANOVA (Tukey), $P = 0.001$) but no difference was observed between ambient and gradual treatments (ANOVA (Tukey), $P = 0.53$).

responses to environmental changes that are significantly more abrupt than those that would occur in nature. □

Methods

Each experimental unit consisted of a single plant growing in an eight-inch pot containing field-collected soil. Plants were initially collected as seed from the field, germinated in the laboratory and added to the experimental units as one-week-old seedlings. After each 15-week growth period, plant shoots were removed and new seedlings were added to each pot. After the first generation, the source of seed was from plants of the previous generation. *B. inermis* is an obligate cross-pollinated plant. Some inbreeding can occur, but this results in seed that is significantly smaller and distorted. Therefore only out-crossed seed was used for the establishment of new generations.

The experiment consisted of 15 experimental units per treatment. The three [CO₂] treatments were imposed using nine environment-controlled growth chambers. Throughout the experiment, other environmental variables were recorded for each chamber, including air temperature, relative humidity and light intensity. Apart from [CO₂], we did not detect any significant difference in any measured environmental variable among treatments. Plants were randomly assigned to a different growth chamber after every five weeks, and the appropriate [CO₂] was re-set. Plants were watered (400 ml) on a weekly basis and fertilized with 250 ml of low-phosphorus, Long-Ashton solution every two weeks.

At harvest, soil was collected using a 10-cm diameter corer. This soil core was then used to trap AMF for determination of AMF species richness and mycorrhizal dependency. This was achieved by mixing the soil with silica sand, adding this mix to a new eight-inch pot, planting a new seedling of *B. inermis*, letting it grow for three months, cutting off the shoots, planting a new seedling and letting it grow for two more months, and finally collecting the resulting AMF spores from a 50-g core sample. A subsample of the spores (200 randomly chosen spores) was used for species identification. The remainder (average = 548 spores) were mixed into a sterile soil/sand medium in a new eight-inch pot that was then used to grow a seedling of *B. inermis* for 12 weeks. Mycorrhizal dependency was calculated as (the biomass of *B. inermis* grown with AMF minus the biomass of plants grown without AMF) divided by the biomass of *B. inermis* grown with AMF. Plant biomass was determined after drying at 60 °C for 36 h. For [P], plant shoot tissue was ashed at 500 °C for 4 h, dissolved in aqua regia and then determined using a mass spectrometer. Root lengths were determined using an image analysis system (WinRhizo, Regent Instruments Inc.). A sub-sample of roots was collected, cleaned of soil, stained with Chlorazol Black E²⁵ and assessed for mycorrhizal colonization²⁶. The total length of fungal hyphae was determined after extraction from a subsample of soil using the filtration method²⁷. The length of living hyphae belonging to each of the different genera was assessed using direct immunofluorescence¹⁴. Finally, as an overall assessment of soil aggregate size distribution we measured the MWD and WSA of the 1–2 mm aggregate size class using a wet-sieving protocol²⁸.

Received 20 July; accepted 9 December 2004; doi:10.1038/nature03268.

- Körner, C. & Bazzaz, F. A. *Carbon Dioxide, Populations, and Communities* (Academic Press, San Diego, 1996).
- Luo, Y. & Mooney, H. A. *Carbon Dioxide and Environmental Stress* (Academic Press, San Diego, 1999).
- Luo, Y. & Reynolds, J. F. Validity of extrapolating field CO₂ experiments to predict carbon sequestration in natural ecosystems. *Ecology* **80**, 1568–1583 (1999).
- IPCC. Climate Change 2001: Synthesis Report. in *A Contribution of Working Groups I, II, and III to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (eds Watson, R. T. & the core writing team) 1–398 (Cambridge Univ. Press, Cambridge, 2001).
- Bazzaz, F. A. The response of natural ecosystems to the rising global CO₂ levels. *Annu. Rev. Ecol. Syst.* **21**, 167–196 (1990).
- Ward, J. K. & Kelly, J. K. Scaling up evolutionary responses to elevated CO₂: lessons from *Arabidopsis*. *Ecol. Lett.* **7**, 427–440 (2004).
- Sinead, C. & Bell, G. Phenotypic consequences of 1,000 generations of selection at elevated CO₂ in a green alga. *Nature* **431**, 566–569.
- Rillig, M. C., Treseder, K. K. & Allen, M. F. in *Mycorrhizal Ecology* (eds van der Heijden, M. G. A. & Sanders, I.) 135–160 (Springer, Berlin, 2002).
- Treseder, K. K. A meta-analysis of mycorrhizal responses to nitrogen, phosphorus, and atmospheric CO₂ in field studies. *New Phytol.* **164**, 347–355 (2004).
- Staddon, P. L. & Fitter, A. H. Does elevated atmospheric carbon dioxide affect arbuscular mycorrhizas? *Trends Ecol. Evol.* **13**, 455–458 (1998).
- Allen, M. F. *Ecology of Mycorrhizae* (Cambridge Univ. Press, Cambridge, 1991).
- Smith, S. E. & Read, D. J. *Mycorrhizal Symbioses* 2nd edn (Academic, London, 1997).
- Klironomos, J. N., Ursic, M., Rillig, M. & Allen, M. F. Inter-specific differences in the response of arbuscular mycorrhizal fungi to *Artemisia tridentata* grown under elevated atmospheric CO₂. *New Phytol.* **138**, 599–605 (1998).
- Treseder, K. K., Egerton-Warburton, L. M., Allen, M. F., Cheng, Y. & Oechel, W. C. Alteration of soil carbon pools and communities of mycorrhizal fungi in chaparral exposed to elevated carbon dioxide. *Ecosystems* **6**, 786–796 (2003).
- Wolf, J., Johnson, N. C., Rowland, D. L. & Reich, P. B. Elevated CO₂ and plant species richness impact arbuscular mycorrhizal fungal spore communities. *New Phytol.* **157**, 579–588 (2003).
- Klironomos, J. N. Variation in plant response to native and exotic arbuscular mycorrhizal fungi. *Ecology* **84**, 2292–2301 (2003).
- Klironomos, J. N. in *Microbial Biosystems: New Frontiers* (eds Bell, C. R., Brylinsky, M. & Johnson-Green, P.) 845–851 (Proc. 8th Int. Symp. on Microbial Ecology, Halifax, 2000).
- Hart, M. M. & Reader, R. J. Taxonomic basis for variation in the colonization strategy of arbuscular mycorrhizal fungi. *New Phytol.* **153**, 335–344 (2002).
- Rillig, M. C., Wright, S. F., Allen, M. F. & Field, C. B. Rise in carbon dioxide changes soil structure. *Nature* **400**, 628 (1999).
- Piotrowski, J. S., Denich, T., Klironomos, J. N., Graham, J. M. & Rillig, M. C. The effects of arbuscular

- mycorrhizae on soil aggregation depend on the interaction between plant and fungal species. *New Phytol.* **164**, 365–373 (2004).
- Van der Heijden, M. G. A. *et al.* Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396**, 69–72 (1998).
- Goverde, M., Van der Heijden, M. G. A., Wiemken, A., Sanders, I. R. & Erhardt, A. Arbuscular mycorrhizal fungi influence life history traits of a lepidopteran herbivore. *Oecologia* **125**, 362–369 (2000).
- Gange, A. C., Brown, V. K. & Alphin, D. M. Multitrophic links between arbuscular mycorrhizal fungi and insect parasitoids. *Ecol. Lett.* **6**, 1051–1055 (2003).
- Rillig, M. C. Arbuscular mycorrhizae and ecosystem processes. *Ecol. Lett.* **7**, 740–754 (2004).
- Brundrett, M. C., Piché, Y. & Peterson, R. L. A new method for observing the morphology of vesicular-arbuscular mycorrhizae. *Can. J. Bot.* **62**, 2128–2134 (1984).
- McGonigle, T. P., Miller, M. H., Evans, D. G., Fairchild, G. L. & Swan, J. A. A new method which gives an objective measure of colonization of roots by vesicular-arbuscular mycorrhizal fungi. *New Phytol.* **115**, 495–501 (1990).
- Miller, R. M., Reinhardt, D. R. & Jastrow, J. D. External hyphal production of vesicular-arbuscular mycorrhizal fungi in pasture and tallgrass prairie communities. *Oecologia* **103**, 17–23 (1995).
- Kemper, W. D. & Rosenau, R. C. in *Methods of Soil Analysis Part 1. Physical and Mineralogical Methods* 2nd edn (ed. Klute, A.) 425 (Agronomy Monograph No. 9, American Society of Agronomy, Madison, 1986).

Acknowledgements We thank T. Crouch, T. Denich, P. Jermaine, P. John, B. Kerin, G. Lefebvre, J. McCune, P. Moutoglou, J. Neville, V. Ortega, Y. Sipsos-Randor, F. Steele, V. Turner, D. Wollad, E. Young and V. Yu for technical assistance. Financial support for this project was provided by a grant to J.N.K. by the Natural Sciences and Engineering Research Council of Canada, to M.E.A. by the US Department of Energy, and to M.C.R. by the US National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.N.K. (jklirono@uoguelph.ca).

Directed aerial descent in canopy ants

Stephen. P. Yanoviak^{1,2}, Robert Dudley^{3,4} & Michael Kaspari^{4,5}

¹University of Texas Medical Branch, 301 University Boulevard, Galveston, Texas 77555, USA

²University of Florida Medical Entomology Laboratory, 200 9th Street SE, Vero Beach, Florida 32962, USA

³Department of Integrative Biology, University of California, Berkeley, California 94720, USA

⁴Smithsonian Tropical Research Institute, PO Box 2072, Balboa, Republic of Panama

⁵Department of Zoology, University of Oklahoma, Norman, Oklahoma 73019, USA

Numerous non-flying arboreal vertebrates use controlled descent (either parachuting or gliding *sensu stricto*^{1,2}) to avoid predation or to locate resources^{3–7}, and directional control during a jump or fall is thought to be an important stage in the evolution of flight^{3,8,9}. Here we show that workers of the neotropical ant *Cephalotes atratus* L. (Hymenoptera: Formicidae) use directed aerial descent to return to their home tree trunk with >80% success during a fall. Videotaped falls reveal that *C. atratus* workers descend abdomen-first through steep glide trajectories at relatively high velocities; a field experiment shows that falling ants use visual cues to locate tree trunks before they hit the forest floor. Smaller workers of *C. atratus*, and smaller species of *Cephalotes* more generally, regain contact with their associated tree trunk over shorter vertical distances than do larger workers. Surveys of common arboreal ants suggest that directed descent occurs in most species of the tribe Cephalotini and arboreal Pseudomyrmecinae, but not in arboreal ponerimorphs or Dolichoderinae. This is the first study to document the mechanics and ecological relevance of this form of locomotion in the Earth's most diverse lineage, the insects.

The tropical forest canopy is home to a large fraction of the Earth's species, and ants are an important part of the associated fauna¹⁰. Canopy ants exhibit numerous adaptations that promote an arboreal lifestyle, such as modified 'sticky' tarsi that allow them to cling to surfaces^{11,12}. Despite these adaptations, large numbers of ants fall from trees in tropical forests, a result of being dislodged by wind or by arboreal mammals or birds^{13,14}. Moreover, *Cephalotes atratus* workers will voluntarily drop off tree trunks when approached by a foreign object¹⁵. This is paradoxical because worker ants lack wings, and the likelihood of any ant rejoining its colony after falling to the forest floor 30 m below is presumably quite low. Here we show that falling *Cephalotes* spp. rarely reach the forest floor. Instead, they control their aerial descent such that they glide back to their host tree trunk, preventing a landing in the understory and facilitating a rapid return to the nest.

Observations of this behaviour in *C. atratus* during previous studies in Panama¹⁶, and recent video recordings in Peru (see Supplementary Videos), showed that falling individuals of *C. atratus* follow a consistent J-shaped path leading back to their associated tree trunk. The trajectory is characterized by three distinct stages: (1) a vertical drop (uncontrolled parachuting); (2) a rapid directional adjustment with body alignment towards the tree; and (3) a steep but directed glide to the trunk. By painting the gaster and hind legs of the ants white, we determined that *C. atratus* workers align their abdomens towards the trunk during the adjustment stage, and that the subsequent glide to the trunk occurs abdomen-first and dorsal side up over the entire transit.

We quantified the consistency of directed descent behaviour by dropping 120 marked *C. atratus* workers from the lower crown branches (about 27 m) of their resident tree near Iquitos, Peru, under windless conditions. A strong majority of the ants (85%) successfully landed on the trunk. This success rate far exceeds that expected for a parachuting insect exhibiting undirected horizontal drift during free fall (8%; G -test, $P < 0.0001$; see 'Null model' in Methods). Upon making contact with the trunk, ants generally adhered there or tumbled downwards a few metres before gaining a foothold, at which point they began to climb the tree. Marked ants typically returned to the same branch, often walking near the original drop point, within 10 min of the fall. This is the first quantification of controlled descent in ants, and the first record of any macroscopic animal consistently gliding backwards to reach a target. Furthermore, a pair of experiments (see Supplementary Information) documents that return success does not depend on familiarity with the colony's host tree, and that the ants align to the target tree regardless of their initial orientation when dropped.

Some ants partly rely on visual stimuli to relocate their nest after foraging^{17,18}. Thus, we predicted that directed descent behaviour in ants depends on their ability to visually locate a tree trunk during a fall. Individual *C. atratus* workers were experimentally blinded by covering the eye with white enamel paint. Control ants and blinded ants also received white paint on the dorsal surface of the head. Blind *C. atratus* workers regained the trunk with significantly lower frequency (average ± 1 s.d. = $10.0 \pm 4.61\%$) than controls ($84.5 \pm 11.52\%$; $P < 0.0001$). The frequency with which blind ants landed on the trunk did not differ from that expected by undirected horizontal motion (G -test, $P > 0.17$; see 'Null model'). Ants used in other experiments occasionally directed their descent towards the beige climbing rope or nearby lianas, suggesting that light-coloured upright linear objects serve as important visual cues for alignment.

Ant gliding performance was analysed from two-dimensional reconstruction of body trajectories obtained within a flight arena (Fig. 1). Of 28 filmed ants, six successfully glided to and landed on a target column within the camera's field of view. Equilibrium glide speeds averaged 4.3 m s^{-1} for these six ants, and the equilibrium glide angle (that is, the angle between the flight path and the horizontal; Fig. 1) averaged 75.0° (range 73.0° – 77.9°), corresponding to a lift:drag ratio of 0.27. Although this value is modest relative

to the shallow glides and high lift:drag ratios exhibited by many gliding vertebrates, such glides by ants are nonetheless impressive given the absence of dedicated lift-generating surfaces and the typical Reynolds number during flight (mean value of 4,110 based on body length). By comparison, a cylinder with length:diameter ratio of four (approximating *C. atratus* workers) will attain a maximum lift:drag ratio of 0.22 at the same Reynolds number (calculated following the resolved-flow analysis in ref. 19). For arboreal ants falling from tall trees, absolute glide angle may be less relevant than the ability to effect stable motion oriented towards the home trunk.

Worker body size (mass) is highly variable within *C. atratus*, and spans two orders of magnitude among species within the genus *Cephalotes* (Fig. 1). We examined the effects of body size on directed descent performance by measuring the vertical distance travelled between the drop point and trunk contact. Glide performance increased (that is, vertical drop distance decreased) with decreasing worker body size in *C. atratus* (Fig. 1). We attribute this to smaller individuals attaining a minimum viable glide velocity ($U_{\min}$; ref. 3) more rapidly than larger individuals during the parachuting phase of the descent.

This size dependence was duplicated interspecifically, further documenting the widespread distribution of directed descent within the myrmicine tribe Cephalotini (Fig. 1). The flattened, flanged morphology typical of Cephalotini is more developed in smaller species^{20,21} and probably enhances their lift:drag ratio. However, dorsoventral flattening does not seem to be a prerequisite for directed descent in ants—several species in the subfamily Pseudomyrmecinae controlled their descent when dropped (Sup-

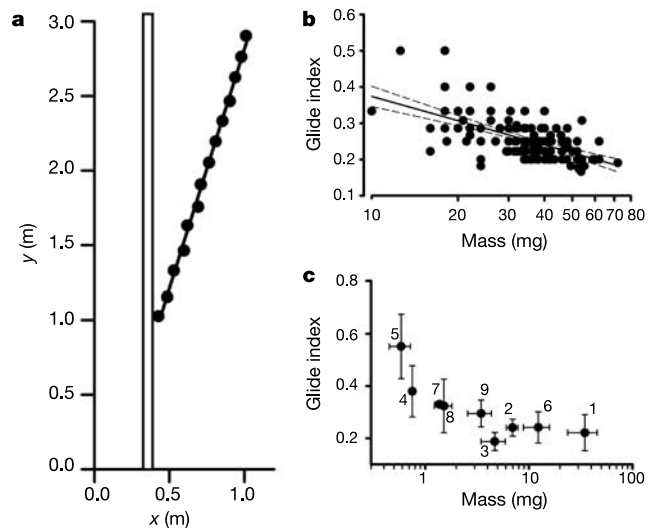


Figure 1 Flight dynamics and performance of *Cephalotes* workers. **a**, Representative aerial trajectory for a *C. atratus* worker gliding to and landing on a vertically aligned white fabric column. Time interval between points is 1/30 s. The regression is given by: $y = 3.22x - 0.41$, $R^2 = 0.996$, $P < 0.001$. Equilibrium glide velocity averaged 4.53 m s^{-1} ; body mass was 0.047 g. **b**, Glide index (horizontal distance travelled per unit vertical drop) as a function of body size in 140 *C. atratus* workers collected from six different colonies (11–45 ants per colony). Some points are hidden by overlap. Mass was measured in the laboratory (59% of points) or estimated in the field (41%) from head width (see Supplementary Information). Solid line is the least square regression. Glide index = $-0.097 \ln(\text{mass}) + 0.596$; $R^2 = 0.355$; $P < 0.0001$; dashed lines are 95% confidence limits. **c**, Average (± 1 s.d.) glide index versus mass for minor workers of nine *Cephalotes* spp. Numbers refer to species listed in Supplementary Table 1. For glide index means, $n = 2$ drops for *C. pallens* (point 7); $n > 2$ for all others. For mass means, $n = 1$ for *C. grandinosus* (point 4); $n > 7$ for all others. Horizontal distance ranged from 2.0 to 2.5 m in **b** and from 1.0 to 2.5 m in **c**.

plementary Table 1); these are more cylindrical ants that lack lateral body flanges.

Directed aerial descent is not exhibited by all arboreal ants. Although it appears to be widespread in the Cephalotini and common in Pseudomyrmecinae, it is rare or absent among tested species in other groups of arboreal ants (Dolichoderinae, ponerimorphs and Formicidae). The following characteristics are consistent with this taxonomic distribution and may be evolutionarily correlated with directed descent behaviour: (1) arboreal nesting; (2) frequent foraging at branch tips (for example, for pollen, nectar and homopteran exudates); (3) relatively costly workers (for example, small colony size and/or heavily armoured individuals); (4) morphology permitting a broad range of abdominal movement^{21,22} (for example, the presence of a post-petiole); (5) good vision and diurnal activity (promoting the use of high-contrast visual cues); and (6) evolutionary origins in flooded forests (where predation pressure from surface-feeding fish²³ may be extreme). Note that *Daceton armigerum* and *Camponotus* cf. *canescens* are taxonomically distinct from the Cephalotini and Pseudomyrmecinae, but both showed controlled descent behaviour (see Supplementary Table 1). *D. armigerum* meets the criteria proposed above²⁴ whereas *C. cf. canescens* does not (at least for lack of a post-petiole). This suggests that further comparative studies throughout the tropics, exploiting the 100+ species that can live in a small patch of forest canopy¹⁴, will help refine our understanding of the ecology and evolution of directed aerial descent in ants. □

Methods

Study areas

Ants were dropped from trees on Barro Colorado Island, Panama (09.15° N, 79.85° W), at the La Selva Biological Station, Costa Rica (10.43° N, 84.02° W), and in western Amazonian forests within 150 km of Iquitos, Peru (03.75° S, 73.25° W; Supplementary Table 2). Vouchers were deposited at the National Museum of Natural History (Washington DC), Smithsonian Tropical Research Institute (Panama), and the Museo de Historia Natural Javier Prado (Peru).

Drop tests

Workers were grasped in the midthoracic region with forceps, held a few centimetres to one side of a branch, and dropped. Each ant was scored according to whether or not it made contact with the trunk before reaching the ground. All drops were initiated during periods of still air, as determined by lack of leaf motion in the surrounding vegetation and by the straight trajectories of dropped inanimate objects. Ants that passed near the tree but landed out of sight (that is, on the side of the tree opposite the observer) were excluded from quantitative analyses (4.5% of the total number of ants dropped; scoring them as misses did not affect the results). Ants were examined for the presence of all appendages before dropping.

Vertical drop distances for performance tests were measured with a string of flags. The string was weighted at one (free hanging) end and secured to the tree at the other, such that it was near or touching the trunk surface over most of its length. Flags were spaced at 1.0-m intervals. Drop points were established in the tree crown, and horizontal distances were measured relative to the attachment point of the string. Vertical drop distances were measured to the point where the ant passed or struck the string, and were estimated to the nearest 0.5 m (0.25 m for drop distances <4 m).

Null model

A falling object will land directly beneath its drop point unless it is acted on by a lateral force (for example, wind) or its aerodynamic profile changes (for example, a parachuting insect moving its appendages). For a drop point on a branch located 2.0 m horizontal from the central axis of the tree trunk, an ant falling in still air but haphazardly moving >2.0 m horizontally relative to the drop point should strike a 1.0 m diameter tree trunk in eight of every 100 falls. The average diameter of trees used for drops was 0.73 m (Supplementary Table 2), thus the null is an overestimate. Incorporating trunk curvature does not substantially increase the null probability at horizontal distances >2.0 m. Ants that did not glide tended to land laterally >2.0 m relative to the drop point (for example, vision experiment), meeting the most fundamental model assumption.

All experiments followed the same general protocol: 10–30 ants from the same colony were dropped from two drop points (branches) separated by 100–180 radians in a tree. Foraging workers from different colonies were used in each replicate. Whenever possible, different trees and source colonies were used for different experiments. Proportional data were arcsine square-root transformed before analysis. The vision experiment was conducted in five replicate trees, and data were analysed with a split-plot analysis of variance (ANOVA) testing for effects of replicates, branches within replicates, and vision manipulation.

Video recordings and glide dynamics

Ants were held transiently at a height of 8 m, and were then dropped approximately 50 cm

in front of a vertically aligned white fabric column (width 36 cm) that visually simulated a flat tree trunk. A tripod-mounted digital video camera (Canon ZR10) was positioned approximately 6 m from and orthogonal to glide trajectories of ants to the fabric column, of which the camera's field of view covered approximately the final three vertical metres. Video frames were downloaded into a microcomputer video editor at 30 frames s⁻¹, and individual frames were then digitized to obtain coordinates of the centroid of the moving ant as a function of time. A horizontal calibration frame located directly behind the plane of glide trajectories was used to determine absolute spatial coordinates. Positional data were evaluated using the program Quicksand²⁵ to obtain translational velocities as a function of time. Equilibrium glides by ants were defined as those glide intervals during which instantaneous translational velocity varied by no more than 5% from the average value for the time interval under consideration. Once such intervals were identified, equilibrium glide angles were determined from the slopes of linear regressions of the associated positional coordinates. A deadweight drop (plastic cylinder 5 cm long by 1.5 cm diameter filled with water) was used to validate the accurate reconstruction (that is, to within 1%) of translational acceleration under gravity alone. Ambient wind was monitored during experiments; no gliding trials were conducted when wind speed exceeded 0.5 m s⁻¹. Body mass of the six ants exhibiting equilibrium glides averaged 47.8 mg; their body length averaged 11.6 mm.

Received 16 September; accepted 8 December 2004; doi:10.1038/nature03254.

- Oliver, J. A. "Gliding" in amphibians and reptiles, with a remark on arboreal adaptation in the lizard, *Anolis carolinensis carolinensis* Voigt. *Am. Nat.* **85**, 171–176 (1951).
- Moffett, M. W. What's "up"? A critical look at the basic terms of canopy biology. *Biotropica* **32**, 569–596 (2000).
- Kingsolver, J. G. & Koehl, M. A. R. Selective factors in the evolution of insect wings. *Annu. Rev. Entomol.* **39**, 425–451 (1994).
- Mori, A. & Hikida, T. Field observations on the social behavior of the flying lizard, *Draco volans sumatranus*, in Borneo. *Copeia* **1994**, 124–130 (1994).
- Jackson, S. M. Glide angle in the genus *Petaurus* and a review of gliding in mammals. *Mamm. Rev.* **30**, 9–30 (1999).
- McCay, M. G. Aerodynamic stability and maneuverability of the gliding frog *Polypedates dennysi*. *J. Exp. Biol.* **204**, 2817–2826 (2001).
- Socha, J. J. Gliding flight in the paradise tree snake. *Nature* **418**, 603–604 (2002).
- Maynard Smith, J. The importance of the nervous system in the evolution of animal flight. *Evolution* **6**, 127–129 (1952).
- Dudley, R. *The Biomechanics of Insect Flight: Form, Function, Evolution* 275–287 (Princeton Univ. Press, Princeton, 2000).
- Stork, N. E., Adis, J. & Didham, R. K. *Canopy Arthropods* (Chapman & Hall, London, 1997).
- Orivel, J., Malherbe, M. C. & Dejean, A. Relationships between pretarsus morphology and arboreal life in ponerine ants of the genus *Pachycondyla* (Formicidae: Ponerinae). *Ann. Entomol. Soc. Am.* **94**, 449–456 (2001).
- Federle, W., Riehle, M., Curtis, A. S. G. & Full, R. J. An integrative study of insect adhesion: mechanics and wet adhesion of pretarsal pads in ants. *Integr. Comp. Biol.* **42**, 1100–1106 (2002).
- Haemig, P. D. Effects of birds on the intensity of ant rain: a terrestrial form of invertebrate drift. *Anim. Behav.* **54**, 89–97 (1997).
- Longino, J. T. & Colwell, R. K. Biodiversity assessment using structured inventory: capturing the ant fauna of a tropical rain forest. *Ecol. Appl.* **7**, 1263–1277 (1997).
- Weber, N. A. The nest of an anomalous colony of the arboreal ant *Cephalotes atratus*. *Psyche (Cambridge)* **64**, 60–69 (1957).
- Yanoviak, S. P. & Kaspari, M. E. Community structure and the habitat template: ant assemblages in the tropical canopy and litter. *Oikos* **89**, 259–266 (2000).
- Wohlgemuth, S., Ronacher, B. & Wehner, R. Ant odometry in the third dimension. *Nature* **411**, 795–798 (2001).
- Hölldobler, B. Canopy orientation: a new kind of orientation in ants. *Science* **210**, 86–88 (1980).
- Ellington, C. P. Aerodynamics and the origin of insect flight. *Adv. Insect Physiol.* **23**, 171–210 (1991).
- Kempf, W. W. A taxonomic study on the ant tribe Cephalotini (Hymenoptera: Formicidae). *Rev. Entomol.* **22**, 1–244 (1951).
- de Andrade, M. L. & Baroni Urbani, C. Diversity and adaptation in the ant genus *Cephalotes*, past and present. *Stuttgarter Beitr. Naturkunde B (Geologie und Paläontologie)* **271**, 1–889 (1999).
- Coyle, F. A. Defensive behavior and associated morphological features in three species of the ant genus *Paracryptocerus*. *Insectes Soc.* **13**, 93–104 (1966).
- Saint-Paul, U. et al. Fish communities in central Amazonian white- and blackwater floodplains. *Environ. Biol. Fishes* **57**, 235–250 (2000).
- Wilson, E. O. Behavior of *Daceton armigerum* (Latreille), with a classification of self-grooming movements in ants. *Bull. Mus. Comp. Zool. Harv.* **127**, 401–422 (1962).
- Walker, J. A. Estimating velocities and accelerations of animal locomotion: a simulation experiment comparing numerical differentiation algorithms. *J. Exp. Biol.* **201**, 981–995 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank B. Fisher, B. Hölldobler, J. T. Longino, S. Combes, G. Byrnes, W. Lamar, J. B. Gonzales and C. Saux for helpful discussions, logistical support, and/or comments on the manuscript. J. T. Longino, W. P. Mackay and P. S. Ward assisted with ant identifications. The Smithsonian Tropical Research Institute, the Panamanian National Authority for the Environment (ANAM), and the Peruvian National Institute of Natural Resources (INRENA) provided permits. This project was supported in part by NSF grants to M.K. and an NIH grant to S. C. Weaver (UTMB).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.K. (mkaspari@ou.edu).

Photoperiodic induction of synchronous flowering near the Equator

Rolf Borchert¹, Susanne S. Renner², Zoraida Calle³, Diego Navarrete⁴, Alan Tye⁵, Laurent Gautier⁶, Rodolphe Spichiger⁶ & Patricio von Hildebrand⁴

¹Biology Division, University of Kansas, Lawrence, Kansas 66045-7534, USA

²Department of Biology, Ludwig Maximilians University, 80638 Munich, Germany

³Center for Research on Sustainable Agriculture CIPAV, A. A. 20591, Cali, Colombia

⁴Fundacion Puerto Rastrojo, A. A. 241438, Bogotá, Colombia

⁵Botany Department, Charles Darwin Research Station, Galapagos, via AP 17-01-3891, Quito, Ecuador

⁶Conservatoire et Jardin botaniques de la Ville de Genève, Case Postale 60, CH-1292 Chambésy/GE, Switzerland

In tropical rainforests, 30–65% of tree species grow at densities of less than one individual per hectare¹. At these low population densities, successful cross-pollination relies on synchronous flowering. In rainforests with low climatic seasonality, photoperiodic control is the only reliable mechanism for inducing synchronous flowering^{2,3}. This poses a problem because there is no variation in day length at the Equator. Here we propose a new mechanism of photoperiodic timekeeping based on the perception of variation in sunrise or sunset time, which explains and predicts the annually repeated, staggered, synchronous and bimodal flowering of many tree species in Amazonian rainforests near the Equator.

Seasonal variation in day length is the only environmental signal that is constant from year to year; it is independent of seasonal and inter-annual variation in climate and is capable of inducing synchronous developmental change in conspecific trees at the same time every year^{2–4}. Photoperiodic control of tropical tree

development has been overlooked, but recent studies have shown that photoperiodic induction of synchronous vegetative bud break or flowering is common in tropical forests around the globe^{3–6}. At tropical latitudes, changes in day length large enough to affect plant development occur only around the equinoxes^{3–7} (Fig. 1a, b). A single annual period of synchronous flowering 1–2 months after the autumn equinox indicates the induction of flowering in 'short-day plants' in response to declining day length⁴ (Fig. 2a; 10° N, 16° S). Flowering of 'long-day plants' and vegetative bud break induced by increasing day length occur after the spring equinox^{3,5–7}. Such flowering periods are six months out of phase between the Northern and Southern Hemispheres^{4,7} (Fig. 2a; compare 10° N and 16° S). In tropical semi-deciduous forests at 10° N and 16° S, photoperiodic control causes the synchronous onset of flowering in all trees of certain species during the same two-week period each year⁴, as indicated by small standard deviations in mean flowering time (Fig. 2a; 10° N).

At 4° N, near Cali, Colombia, flowering of more than 100 *Montanoa quadrangularis* trees was observed during the same two-week period over three consecutive years⁸ (Fig. 2a; 4° N). This raised the question of how such synchronized flowering near the Equator could be induced by day length changes of <30 min year⁻¹ (Fig. 1a) or <4 min over 20 days (Fig. 1b). At the Equator, where day length is constant throughout the year (Fig. 1a, red line), the times of sunrise and sunset (Fig. 1c, e, red lines) vary by 30 min over the course of the year (<http://www.geocities.com/jllammi/>) and therefore might act as a celestial flowering signal. Annual variation of sunrise time at the Equator corresponds to the difference between apparent solar time measured by a sundial and mean time measured by chronometers⁹ ('equation of time'; Fig. 1c, red line, right y-axis). In contrast to seasonal variation in day length (Fig. 1a), sunrise time has two annual maxima and minima, corresponding to the largest delays or advances of solar time relative to mean time (Fig. 1c, red line). The asymmetric, bimodal shape of this curve is generated by the combination of two properties of the Earth's orbit around the Sun: its elliptic (rather than circular) shape and the tilt of the Earth's axis relative to its orbit⁹ (see Supplementary Discussion 1). With increasing latitude, the asymmetry of the

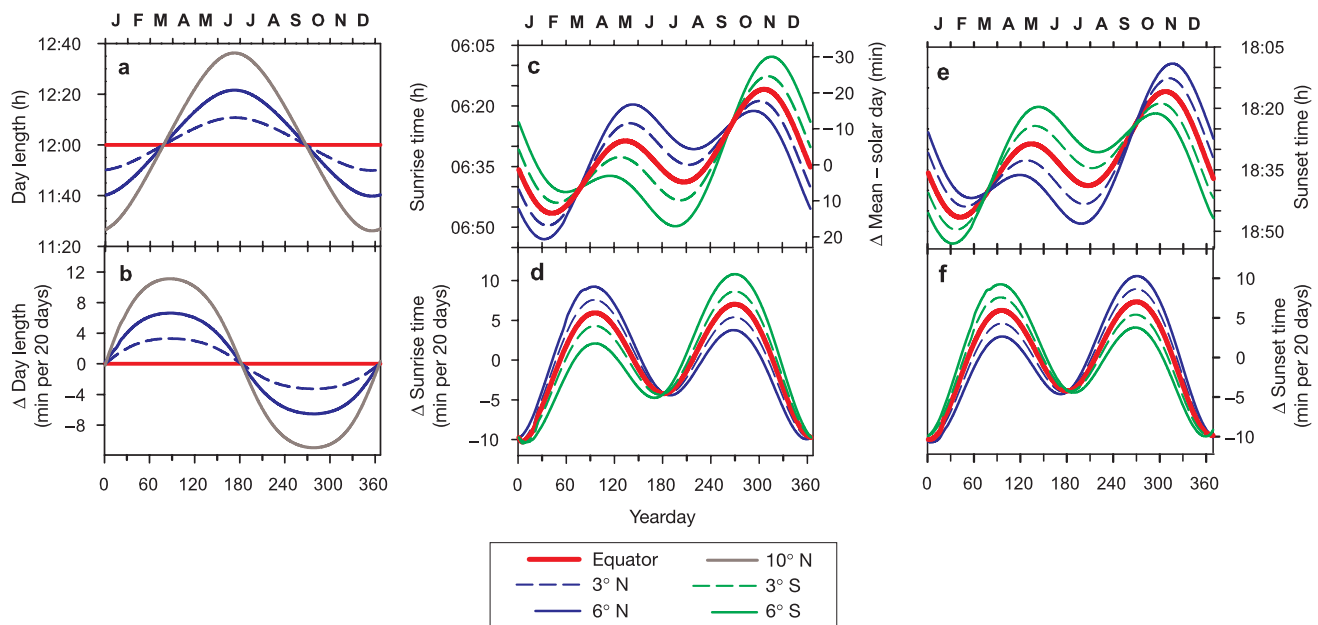


Figure 1 Seasonal variation in day length, sunrise and sunset time near the Equator. Day length (a), sunrise (c) and sunset (e) times obtained from the internet (<http://www.geocities.com/jllammi/>) were used to calculate running 20-day differences (b, d, f).

The equation of time equals seasonal variation in sunrise time at the Equator (c, right y-axis, red line).

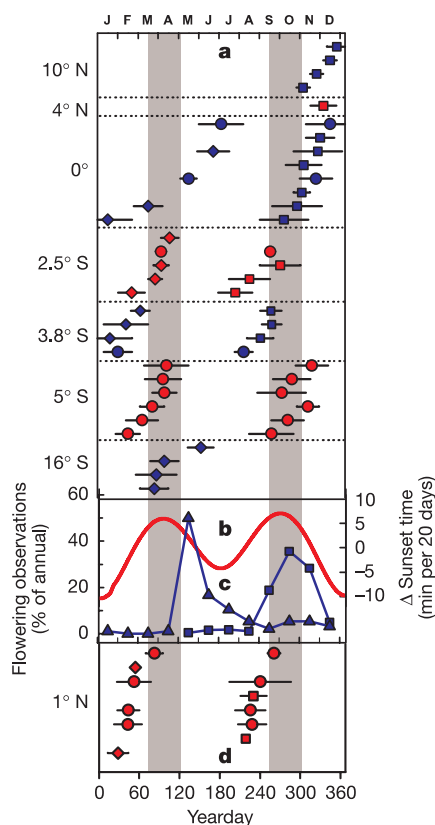


Figure 2 Synchronous, bimodal flowering and bud break in tropical forests at low latitudes. **a**, Mean times of synchronous flowering in 41 representative tree species observed in seven tropical forests ranging from 10°N to 16°S. For each forest site, the duration of the observation period and the author name for any unpublished field observations are given. Species names (for each site, from left to right and top to bottom; diamonds indicate spring-flowering species, squares indicate autumn-flowering species, circles indicate species flowering in the spring or autumn) are followed by the number of trees used to calculate mean flowering time and standard deviation (error bars). 10°N: Tropical semi-deciduous forest in Guanacaste, Costa Rica⁴ (3 yr); *Cochlospermum vitifolium* (>100), *Cordia alliodora* (>100), *Calycophyllum candidissimum* (>100), *Bixa orellana* (10). 4°N: Tropical montane forest in El Dovio, Valle de Cauca, Colombia⁸ (2 yr); *Montanoa quadrangularis* (113). 0°: Rainforest at Parque Nacional Chiribiquete, Caquetá, Colombia (D.N., 3.5 yr); *Clusia chiribiquetensis* (143), *Peltogyne catingae* (12), *Pouteria* sp. (26), *Manilkara bidentata* (19), *Iryanthera obovata* (15), *Mollia speciosa* (13), *Clathrotropis macrocarpa* (287), *Erisma laurifolium* (64), *Micropholis* sp. (13), *Protium guianense* (11), *Iryanthera tricornis* (13). 2.5°S: Rainforest at Manaus, Amazonia, Brazil (S.S.R., 3 yr); *Miconia longispicata* (10), *M. pyriformis* (8), *M. elaeagnoides* (12), *M. regelii* (9), *M. burchellii* (31), *M. egensis* (8), *M. myriantha* (7), *M. dispar* (9). 3.8°S: Rainforest at Parque Nacional Amacayacu, Amazonas, Colombia (A.T., 2.2 yr); *Miconia* cf. *cayumbensis* (11), *Randia ruiziana* (4), *Swartzia auriculata* (7), *Coussarea paniculata* (7), *Henriettea stellaris* (9), *Psychotria* cf. *capitata* (6), *Miconia sprucei* (3). 5°S: Rainforest at Arborescencia Jenaro Herrera, Requena, Loreto, Peru^{15,16} (8 yr); *Caryocar glabrum* (37), *Quararibea ochrocalyx* (17), *Helicostylis scabra* (9), *Swartzia cuspidata* (6), *Eschweilera tessmannii* (22), *Diclinanona tessmannii* (24). 16°S: Tropical semi-deciduous forest at Lomerio, Ñuflo de Chavez, Santa Cruz, Bolivia^{4,17} (3 yr); *Vochysia mapiensis* (10), *Calycophyllum multiflorum* (99), *Pterogyne nitens* (87), *Centrolobium microchaete* (84). **b**, Seasonal variation in sunset time. Vertical grey bars indicate periods of maximum advance in sunset time. **c**, Flowering of eight spring-flowering (triangles) and 27 autumn-flowering species (squares) in the equatorial rainforest at Parque Nacional Chiribiquete, Caquetá, Colombia (D.N.). **d**, Vegetative bud break observed over ten years in individual trees of deciduous or leaf-exchanging species in the Singapore Botanical Garden at 1°N¹⁰. Four species exchanged leaves twice a year (circles). Top to bottom: *Couroupita guianensis*, *Parkia javanica*, *Ficus variegata*, *Lecythis* sp., *Peltophorum pterocarpum*, *Terminalia catappa*, *Cedrela glaziovii*, *Kigelia pinnata*.

curve increases as the effects of the equation of time combine with progressively larger seasonal variation in day length (Fig. 1c).

If induced by bimodal variation in sunrise or sunset time, synchronous flowering near the Equator should occur after both equinoxes, when these variables change fastest (Fig. 1d). Flowering synchrony observed in four 2–8-yr phenological surveys of thousands of trees in Amazonian rainforests within 5° of the Equator is generally less precise than at higher latitudes; this is probably because of the marginal photoperiodic signal. Species with synchronous flowering were therefore identified as those for which flowering occurred at the same time in different years and >75% of flowering observations were made within three-month periods following several months with low or no flowering activity (Fig. 2c; see Supplementary Information). Most of these species flowered during the predicted periods (Fig. 2a–c, grey bars). Two annual fruiting periods have also been observed in other species (A.T., unpublished observation). The striking coincidence between the two annual maxima of advancing sunrise/sunset time (at the equinoxes) and the onset of two annual flowering periods in equatorial rainforests provides strong evidence for the importance of advancing sunrise/sunset time in photoperiodic timekeeping. There is no other bimodal environmental signal in these forests that occurs at precisely the same time each year. No climatic cue, such as rainfall or temperature, can explain an identical sequence of species' flowering times year after year. In South America, bimodal flowering caused by bimodal rainfall occurs only in seasonally dry forests at ~4°N, and its timing varies from year to year with seasonal rainfall (see Supplementary Discussion 2).

The highly synchronous, bimodal vegetative bud break observed over ten years in deciduous and leaf-exchanging tree species in the equatorial climate of Singapore at 1°N is also indicative of photoperiodic control¹⁰ (Fig. 2d). In some of these species, increasing day length is known to induce synchronous vegetative bud break after the spring equinox in mid-latitude tropical forests ('spring flushing'³; Fig. 2d, *Parkia*, *Cedrela*). Synchronous bud break soon after the summer and winter solstices suggests that seasonal delays in sunrise or sunset time, rather than changing day length, induce the breaking of bud dormancy (Figs 1c and 2d). In contrast, vegetative development of *Bombax malabaricum* is asynchronous in Singapore, but synchronous at 5°N and 5°S¹⁰, indicating that perception of changes in sunset time may be less sensitive in this species. Similarly, at higher tropical latitudes, bud break of *Bursera* is induced by increasing day length, but on the Galapagos Islands (0°) it is triggered by the first rains after seasonal drought (A.T., unpublished observation).

Plants use different pigment systems to perceive the large changes in light intensity at dawn and dusk^{2,11}. They measure changes in length of the solar day by relating the signals at dawn and dusk to the phases of their endogenous circadian clock. This process is not well understood^{2,11,12}, but it seems likely that changes in sunrise or sunset time could also be measured by this mechanism. The minimum change in day length required to induce flowering is not known. Experiments with the short-day plant rice (*Oryza sativa*) at 2°N suggest that a 15-min change in day length accumulated over a few weeks may be sufficient¹³. To assess the potentially inductive changes in day length versus sunrise and sunset time near the Equator, we calculated the cumulative changes of these variables over an assumed 20-day induction period (Fig. 1b, d, f; see Supplementary Table). Around the autumn equinox, the advance in sunrise time decreases with increasing latitude, but the advance in sunset time increases (compare Fig. 1d and f: 6°N and 3°N at yearday 270, 6°N and 3°S at yearday 90). These observations suggest that at latitudes between 6°N and S, cumulative changes in sunset time rather than the much smaller changes in sunrise time or day length induce autumn-flowering of short-day plants (Fig. 2a; 4°N to 5°S). Correspondingly, at and near the Equator, flowering around the autumn equinox appears to be more synchronized than around the

spring equinox (Fig. 2a; 0° and 2° S), when the advances in sunset time are slightly smaller (7 and 7.6 min versus 5.9 and 5.4 min). Implicitly, cumulative changes in sunset time of 5–7 min over 20 days are sufficient to induce flowering at the Equator.

In view of the small maximum changes in sunset time near the Equator (5–7 min over 20 days) it seems likely that signals sufficiently large to trigger flowering or vegetative bud break occur only during periods of maximum change (Fig. 1 d, f), that is, around the equinoxes (Fig. 2, grey bars) or solstices (Fig. 2d). If so, the observed staggered flowering times (Fig. 2a) could be primarily attributable to differences in the duration of flower development induced by perception of the photoperiodic signal. Flowering during an inductive period might indicate rapid flower emergence from resting buds, as in several *Miconia* species (Fig. 2a; 2.5–3.8° S). In other species, trees may flower months after the inductive period because the flowering signal causes the vegetative shoot apex to change into a large, branched inflorescence supporting many flowers⁴ (Fig. 2a; 4° N *Montanoa*). It remains to be explained why some species flower only during or shortly before one of the two induction periods and why many trees do not flower every year.

This is the first study that both confirms synchronous flowering in rainforest tree species near the Equator and proposes a timing mechanism. Synchronous flowering at the same time each year has long been noticed in Amazonian trees such as *Miconia* (Fig. 2a; 2.5° S), but studies of the phenomenon focused on the evolutionary consequences of staggered synchronous fruiting, for example, for frugivorous birds¹⁴. Photoperiodic control of vegetative development and flowering in tropical trees evolved in response to different adaptive pressures^{3,5,6}. In tropical rainforests with an equitable climate, it may have evolved in response to the need to synchronize flowering to achieve cross-pollination in spite of low population densities. □

Received 10 September; accepted 8 December 2004; doi:10.1038/nature03259.

- Condit, R. *et al.* Tropical tree diversity: Results from a worldwide network of large plots. *Biol. Skr. Dan. Vid. Selsk. (Copenhagen)* (in the press).
- Thomas, B. & Vince-Prue, D. *Photoperiodism in Plants* (Academic, San Diego, 1997).
- Rivera, G. *et al.* Increasing day length induces spring flushing of tropical dry forest trees in the absence of rain. *Trees – Struct. Funct.* **16**, 445–456 (2002).
- Rivera, G. & Borchert, R. Induction of flowering in tropical trees by a 30-min reduction in photoperiod: evidence from field observations and herbarium specimens. *Tree Physiol.* **21**, 201–212 (2001).
- Borchert, R. & Rivera, G. Photoperiodic control of seasonal development in tropical stem-succulent trees. *Tree Physiol.* **21**, 213–221 (2001).
- Borchert, R., Meyer, S. A., Felger, R. S. & Porter-Bolland, L. Environmental control of flowering periodicity in Costa Rican and Mexican tropical dry forests. *Global Ecol. Biogeogr.* **13**, 409–425 (2004).
- Borchert, R. Environmental control of tropical tree phenology (http://www.biology.ku.edu/tropical_tree_phenology/) (2004).
- Calle, Z. Temporal variation in the reproductive phenology of *Montanoa quadrangularis* in the Andes of Colombia: declining photoperiod as a likely environmental trigger. *Biotropica* **34**, 621–622 (2002).
- National Maritime Museum. *Timekeeping: the Equation of Time* (<http://www.nmm.ac.uk/site/request/setTemplate:singlecontent/contentTypeA/conWebDoc/contentId/351>).
- Holtum, R. E. Periodic leaf-exchange and flowering of trees in Singapore (II). *Gardens Bull. S.S.* **11**, 119–175 (1940).
- Imaizumi, T., Tran, H. G., Swartz, T. E., Briggs, W. R. & Kay, S. A. FKF1 is essential for photoperiodic-specific light signaling in *Arabidopsis*. *Nature* **426**, 302–306 (2004).
- Valverde, F. *et al.* Photoreceptor regulation of CONSTANS protein in photoperiodic flowering. *Science* **303**, 1003–1006 (2004).
- Dore, J. Response of rice to small differences in length of day. *Nature* **183**, 413–414 (1959).
- Snow, D. W. A possible selective factor in the evolution of fruiting seasons in tropical forest. *Oikos* **15**, 274–281 (1965).
- Gautier, L. *Reproduction de la Strate Arborée d'une Forêt Ombrophile d'Amazonie Péruvienne* M.S. thesis, Univ. Genève (1985).
- Gautier, L. & Spichiger, R. Ritmos de reproducción en el estrato arbóreo del Arboretum Jenaro Herrera (provincia de Requena, departamento de Loreto, Perú). Contribución al estudio de la flora y de la vegetación de la Amazonia peruana. *X. Candollea* **41**, 193–207 (1986).
- Justiniano, M. J. & Fredericksen, T. S. Phenology of tree species in Bolivian dry forests. *Biotropica* **32**, 276–286 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.B. (borchert@ku.edu).

Gene transfer to plants by diverse species of bacteria

Wim Broothaerts^{*†}, Heidi J. Mitchell[†], Brian Weir[†], Sarah Kaines^{*}, Leon M. A. Smith, Wei Yang, Jorge E. Mayer^{*}, Carolina Roa-Rodriguez^{*} & Richard A. Jefferson

CAMBIA (An Affiliated Research Centre of Charles Sturt University), G.P.O. Box 3200, Canberra, ACT 2601, Australia

^{*} Present addresses: European Commission, Joint Research Centre, Institute for Reference Materials and Measurement, Retieseweg 111, B-2440 Geel, Belgium (W.B.); Research School of Biological Sciences, Australian National University, ACT 2601, Australia (S.K.); Campus Technologies Freiburg, University of Freiburg, Stefan Meier Str. 8, Freiburg D-79104, Germany (J.E.M.); RegNet, Research School of Social Sciences, Australian National University, ACT 2601, Australia (C.R.-R.)

[†] These authors contributed equally to this work

Agrobacterium is widely considered to be the only bacterial genus capable of transferring genes to plants. When suitably modified, *Agrobacterium* has become the most effective vector for gene transfer in plant biotechnology¹. However, the complexity of the patent landscape² has created both real and perceived obstacles to the effective use of this technology for agricultural improvements by many public and private organizations worldwide. Here we show that several species of bacteria outside the *Agrobacterium* genus can be modified to mediate gene transfer to a number of diverse plants. These plant-associated symbiotic bacteria were made competent for gene transfer by acquisition of both a disarmed Ti plasmid and a suitable binary vector. This alternative to *Agrobacterium*-mediated technology for crop improvement, in addition to affording a versatile 'open source' platform for plant biotechnology, may lead to new uses of natural bacteria–plant interactions to achieve plant transformation.

Agrobacterium tumefaciens is a ubiquitous soil bacterium that induces galls on plants. The discovery that this gall formation is due to integration into the plant genome of bacterial DNA (T-DNA) laid the foundations for plant biotechnology³. The T-DNA is part of the ~200 kb Ti (tumour-inducing) plasmid, which also encodes functions for Ti plasmid conjugation, opine metabolism and the initiation, transfer and processing of the T-DNA^{4,5}. Before the discovery of the Ti plasmid, gall-inducing ability was shown to be transferable to non-virulent *Agrobacteria* and to *Rhizobium leguminosarum*⁶. Ti plasmid transfer to *Rhizobium trifolii* and *Phyllobacterium myrsinacearum* resulted in strains that caused galls on some plants^{7,8}, but a *Sinorhizobium meliloti* strain containing a Ti plasmid was not tumorigenic⁹. Although these experiments showed that close relatives of *Agrobacterium* could harbour the Ti plasmid, no direct molecular evidence of gene transfer to plants by these bacteria was reported, leaving open the possibility that gall formation may have resulted from hormonal perturbations in the host plant unrelated to DNA transfer¹⁰. Indeed, a disarmed Ti plasmid and binary vector were introduced into a bacterial isolate apparently related to *Phyllobacterium* spp. for the purpose of tobacco inoculation¹¹, and although galls resulted from production of auxin by *Phyllobacterium*, these galls were morphologically different from those produced by an *Agrobacterium*-transformed plant through gene transfer; moreover, evidence of gene transfer was sought but not found. Accordingly, the scientific community has focused on *Agrobacterium* as a vehicle for gene transfer; the vast majority of patent claims regarding biological plant transformation explicitly refer to *Agrobacterium*². A recent proposal suggesting that *A. tumefaciens* be reclassified as *Rhizobium radiobacter* has been widely disputed¹², although *Agrobacterium* is clearly closely related to *Rhizobium*. However there is little doubt that *Agrobacterium*, *Sinorhizobium* and *Mesorhizobium* are in distinct phylogenetic clades and their genomic organization differs considerably^{4,13}.

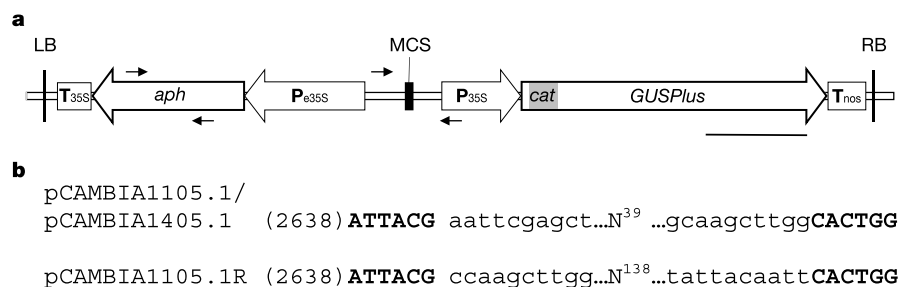


Figure 1 Binary vectors used for plant transformation. **a**, T-DNA map for pCambia vectors used, containing the hygromycin resistance gene (*aph*) for plant selection and the *GUSPlus* gene as a reporter to investigate gene transfer. P35S and P_{35S}, promoter and enhancer + promoter from CaMV35S, respectively; Tnos and T35S, terminators from *A. tumefaciens* nopaline synthase gene and CaMV35S, respectively; cat, catalase intron;

LB and RB, left and right T-DNA border sequence. Arrows show the position of primers used for analysis of T-DNA integration; bar indicates position of probe used for Southern blot analysis. **b**, Multiple cloning site sequences in pCambia1105.1 and pCambia1405.1 (used in *A. tumefaciens*) and pCambia1105.1R (exclusively used in non-*Agrobacterium* bacterial species).

To see whether non-*Agrobacterium* species could be made competent for gene transfer to plants, we introduced the disabled Ti plasmid (pEHA105) from a hypervirulent *Agrobacterium* strain¹⁴ into several different species of bacteria. To facilitate transfer of this large plasmid, we integrated the origin of transfer (*oriT*) of a broad-host range IncP plasmid into the Ti plasmid of EHA105 at two different locations (pTiWB1, pTiWB3). We then mobilized the modified Ti plasmids into *Rhizobium* sp. NGR234, an exceptionally broad-host range *Rhizobium* species that nodulates over 100 different plants¹⁵; the alfalfa symbiont *Sinorhizobium meliloti*; and *Mesorhizobium loti*, a representative of a different family (Phyllobacteriaceae). To assay gene transfer to plants, we introduced the binary vector pCambia1105.1R into the non-*Agrobacterium* bacteria (Fig. 1a) and pCambia1105.1 or pCambia1405.1 into *Agrobacterium*. To confirm the provenance of insertion events, these vectors were constructed with unique T-DNA 'tags' giving rise to polymerase chain reaction (PCR) products of different size. The 'R' version (pCambia1105.1R) was reserved for use only in rhizobia (Fig. 1b) and was never introduced into *Agrobacterium*. We developed additional primers to check the genotype of engineered strains by PCR (Fig. 2) and confirmed that the strains were free of contaminating *Agrobacterium* by selective plating.

Gene transfer to tobacco was investigated by co-cultivating leaf discs with bacterial strains and assaying for β -glucuronidase (GUS) activity (Fig. 3). *GUSPlus*, a synthetic *gusA* gene based on the sequence from *Staphylococcus* sp., produces a GUS protein detectable at up to tenfold higher sensitivity compared to *E. coli* GUS¹⁶. The intron inserted into the *GUSPlus* coding sequence prevents expression in bacterial cells, thus restricting activity to plant cells that had incorporated and expressed the transgene. All strains tested gave rise to GUS-expressing foci on the leaves and leaf margins, but these foci varied considerably in number and location (Fig. 3a).

GUS activity was not observed in control experiments using *S. meliloti* or *Rhizobium* sp. that carried the binary vector but no Ti plasmid, suggesting that gene transfer was mediated through the *vir* gene machinery of pTiWB1 or pTiWB3. The number of GUS-expressing foci observed for tobacco following co-cultivation with various bacterial strains is compiled in Table 1. For non-*Agrobacterium* strains, the developmental stage of the leaves affected gene transfer considerably, with young leaves yielding an average of >10 times more GUS-expressing foci compared with older, just fully expanded leaves over several experiments. No difference was observed between the two leaf stages for *Agrobacterium* transformation. Longer co-culture times also increased gene transfer from non-*Agrobacterium* strains; this is perhaps associated with their slower growth.

To confirm that the transferred T-DNA integrated into the genome, plants were regenerated from transformation events using three non-*Agrobacterium* species (Table 2). Regenerated plants showed GUS activity in their leaves (Fig. 3b), and PCR amplification of the 'provenance-tagged' site confirmed that the T-DNA originated from the binary vector exclusively introduced into non-*Agrobacterium* strains (Fig. 3c). Eleven plants were analysed by Southern blotting, showing stable integration of T-DNA at 1–3 sites per plant (Fig. 3d). Sequences of T-DNA insertion site(s) were determined for several plants, revealing integration into independent loci (Fig. 3e), some of them matching known tobacco sequences. This suggests that a number of diverse plant-associated bacteria, when harbouring a Ti plasmid and binary vector, are able to transfer T-DNA to plants.

The bacterial strains were also tested for their ability to transform other plant species. Rice calli that were co-cultivated with *S. meliloti* (containing pTiWB3 and pCambia1405.1) revealed GUS activity in four out of 687 calli (0.6%). Subsequently, using pTiWB3 and

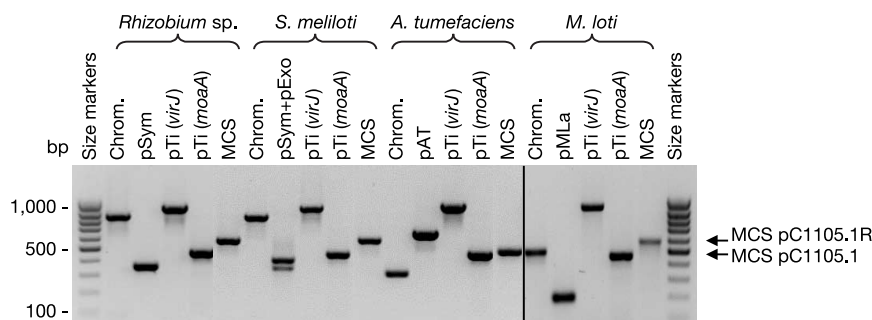


Figure 2 Genotyping of gene transfer-competent bacterial strains containing pTiWB3 by selective PCR amplification of chromosomal and plasmid markers (see Supplementary

Table S1), and the differently sized MCS in pCambia1105.1R (pC1105.1R) and pCambia1105.1 (pC1105.1) vectors (arrows).

pCAMBIA1105.1R, one transformed rice plant (from a total of 695 calli) regenerated and rooted (Fig. 3b, c). Analysis of the single T-DNA integration site in this rice plant (Fig. 3d) indicated that the T-DNA had integrated into rice chromosome 11 (BAC clone OSJNBa0081F16; Fig. 3e). In contrast, in a typical rice transformation experiment using *Agrobacterium*, 50–80% of rice calli co-cultivated for 2 days show GUS-expressing foci and 20% of all calli regenerate transformed shoots.

Arabidopsis transformation¹⁷ following floral dip in a *S. meliloti* suspension resulted in six hygromycin-resistant seedlings from approximately 70,000 seeds, representing 5–10% of the efficiency of *A. tumefaciens* using the same protocol. One plant was also obtained from *Rhizobium*-mediated transformation of *Arabidopsis*. All seven plants expressed GUS in their leaves (Fig. 3b) and contained the T-DNA from pCAMBIA1105.1R (Fig. 3c). Southern blot analysis of two of the plants transformed by *S. meliloti* indicated T-DNA integration at one and three genomic loci respectively (Fig. 3d). The single T-DNA insertion in plant number 4 showed a perfect match to the *A. thaliana* protein phosphatase 2C gene on chromosome I (Fig. 3e). Sequencing of the T-DNA insertion site for plant 5 showed a match at the left and right border with a site in *Arabidopsis* chromosome I (BAC clones T23G18 and T6D22), with a

Table 1 Gene transfer to tobacco by diverse species of bacteria

Species	Ti plasmid*	Number of leaf discs† assayed (number of experiments)	Average number of GUS foci per leaf disc ± s.d.‡
<i>S. meliloti</i>	pTiWB1	40 (4)	49 ± 50
	pTiWB3	69 (7)	53 ± 55
<i>Rhizobium</i> sp. NGR234	pTiWB1	40 (4)	9 ± 12
	pTiWB3	59 (6)	8 ± 15
<i>M. loti</i>	pTiWB1	20 (2)	3 ± 5

*In these experiments, the Ti plasmids are not being compared for gene transfer efficiency as they were not always assayed in parallel. In *Agrobacterium*, similar tobacco transformation efficiencies were observed with either the wild-type Ti plasmid or pTiWB1 or pTiWB3 (data not shown).

†For the data in this table, leaf discs were cut from young (not fully expanded) leaves.

‡Similar experiments with *A. tumefaciens* yielded a typical transformation efficiency of 200–400 GUS foci per leaf disc.

20 bp deletion at the insertion site. The T-DNA insertion site for plant 6 was on chromosome III (*Arabidopsis* BAC clone F16B3). In a subsequent floral dip experiment with *S. meliloti* in which the infiltration medium was modified, two transformed plants were obtained from approximately 4,000 seeds, a fourfold improvement over previous conditions.

These results confirm that all three non-*Agrobacterium* species

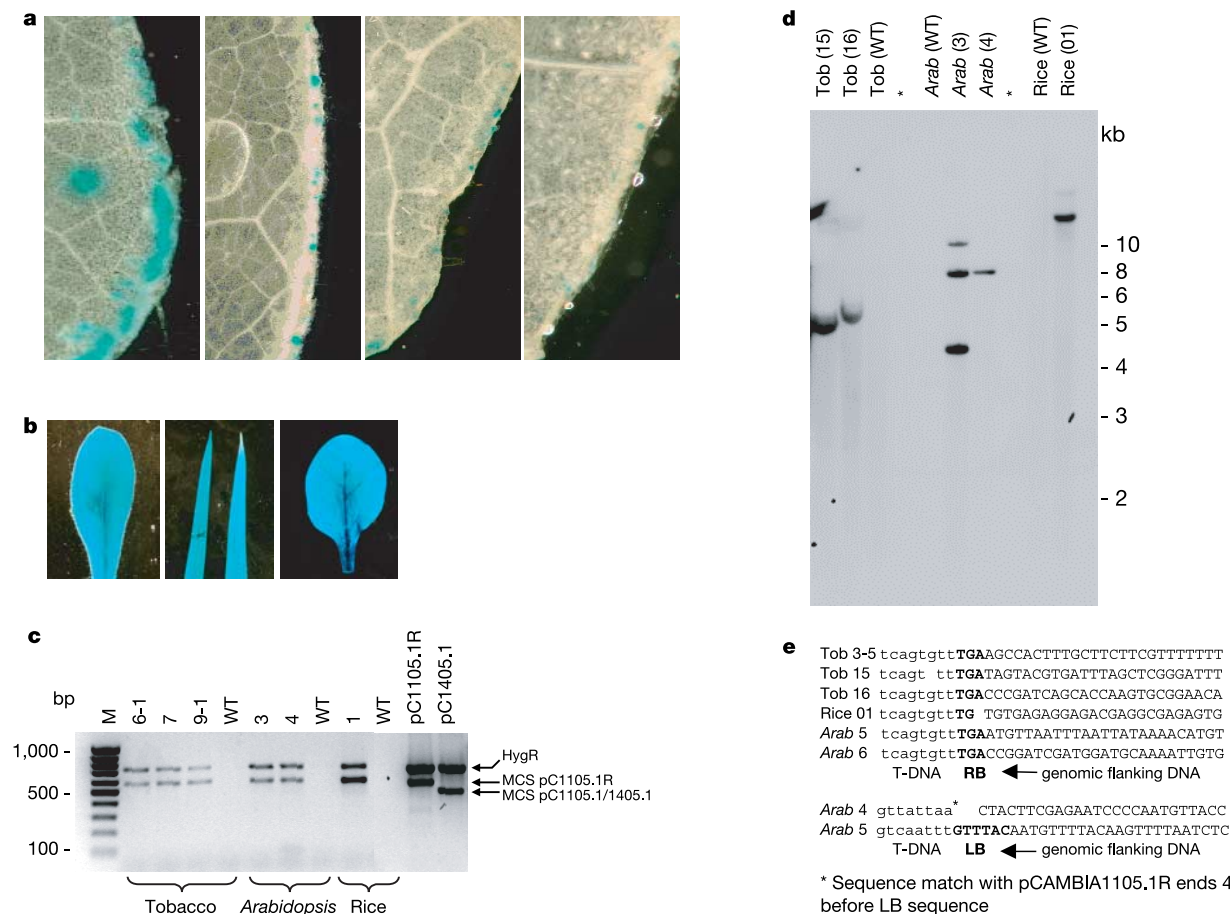


Figure 3 Gene transfer to plants by diverse species of bacteria. **a**, Tobacco leaf discs stained for GUS activity immediately after co-cultivation with (left to right) *A. tumefaciens*, *Rhizobium* sp. NGR234, *S. meliloti* or *M. loti*. **b**, GUS activity in leaves of tobacco (left), rice (middle) and *Arabidopsis* (right), regenerated following transformation with *Rhizobium* sp. NGR234 (tobacco) or *S. meliloti* (rice and *Arabidopsis*). **c**, Duplex PCR analysis for the hygromycin resistance gene (HygR, upper band) and the T-DNA MCS (lower band) in plants transformed through *Rhizobium* sp. NGR234 (tobacco) or *S. meliloti* (*Arabidopsis* and rice). WT, untransformed plant; plasmid controls pCAMBIA1105.1R (pC1105.1R), pCAMBIA1405.1 (pC1405.1); numbers refer to different transformants. **d**, Southern blot

analysis of transgenic tobacco (Tob), *Arabidopsis* (Arab) and rice plants and their respective untransformed controls (WT). Plant number indicated in parentheses. Transgenic plants shown here resulted from *S. meliloti*-mediated transformation. Asterisks used to denote an empty lane. **e**, Sequences of the T-DNA RB insertion site in three tobacco plants and a rice plant, and of the RB or LB insertion sites in three *Arabidopsis* plants after *S. meliloti*-mediated transformation. Lower case letters indicate sequence of T-DNA; bold letters indicate the border sequence of the T-DNA (RB, LB); upper case letters indicate flanking DNA sequences.

Table 2 Recovery of transgenic tobacco plants following gene transfer mediated by bacteria

Bacterial strain	Number of leaf discs (number of experiments)	Number of transformed plants*	Transformation frequency† (%)
<i>S. meliloti</i> pTiWB3	116 (7)	33	28
<i>S. meliloti</i> pTiWB1	22 (2)	8	36
<i>Rhizobium</i> sp. pTiWB3	148 (11)	13	9
<i>M. loti</i> pTiWB1	30 (3)	4‡	13

*T-DNA confirmed by PCR.

†Number of transformed plants obtained per 100 leaf discs. Similar experiments with *A. tumefaciens* yielded a typical transformation frequency of 95–100%.

‡Un-rooted shoots.

tested here can transform plants. Of these, at least *S. meliloti* is competent to transfer genes into both dicot and monocot plants and into a range of tissues, including leaf tissue, undifferentiated calli and immature ovules.

Genes not resident on the Ti plasmid, but located on the *Agrobacterium* chromosomes, have been described as essential for virulence¹. Gene transfer is, however, only a component of the more complex virulence phenotype, and may not require these additional loci. Our observations suggest that if there are gene functions necessary for gene transfer that are not encoded by the Ti plasmid, they must have equivalents or homologues¹³ in multiple rhizobial species. However it is possible that the small number of *vir*-related genes on the Ti plasmid are sufficient to confer gene transfer competence to any bacterium.

Our results suggest a mechanism by which natural, bacterially mediated horizontal gene transfer may have contributed to the composition of plant genomes. There are many genes of presumed bacterial origin in the genomes of plants and other eukaryotes^{18–21}. In a simple case, an apparently ancient integration of T-DNA-derived *rol* genes into the wild-type tobacco genome has been described²². However, it is possible that more substantial gene transfer has occurred over an evolutionary timescale. This could have been mediated by Ti-encoded plant gene transfer or a similar conjugation pathway, but in a manner analogous to the Hfr phenomenon in bacterial mating²³ in which larger bacterial genome segments are transferred to recipient cells.

Many important crops, including species particularly relevant to the developing world, are difficult to transform by *Agrobacterium*, difficult to regenerate from a susceptible explant, or prone to unsuitable genetic and phenotypic variation using current methodology. *Agrobacterium* has evolved as a plant pathogen, presumably with constraints on its specific binding and infection characteristics, and with co-evolved plant responses to pathogenicity. Other plant-associated bacteria, including *Sinorhizobium*, *Rhizobium* and *Mesorhizobium*, interact with different cell types and stages, and as symbionts or benign endo- and epiphytes can be expected to evoke different plant responses. We may benefit from this broader range of interactions to facilitate gene transfer to previously intractable plant cell types, explants or species.

Optimization of this gene transfer technology may require changes to the genetics and/or the physiology of the host strain and the target plant species. Introducing *Agrobacterium* chromosomal genes into new bacterial species could enhance the induction or expression of *vir* genes²⁴. Matching bacterial and plant genotypes or modifying culture conditions may also prove extremely effective. The tenfold enhancement in tobacco transformation using younger explants suggests many avenues for further improvements. Similarly, changes to the infiltration medium for *Arabidopsis* floral dip with *Sinorhizobium* increased the transformation frequency to almost half of the typical frequency using *Agrobacterium*, already making this a viable alternative technology.

We have developed gene transfer competent *Rhizobium*, *Sinorhizobium* and *Mesorhizobium* strains, representing two different

families of bacteria. This approach, based on acquisition of a disarmed Ti plasmid and a binary vector, is unlikely to be limited to these species of bacteria. In contrast to the complex licensing landscape for *Agrobacterium* methodology², this alternative technology is available to the international community in a 'protected technology commons' optimized and improved as a BioForge project (<http://www.bioforge.net>), and accessible under a BIOS (Biological Innovation for Open Society, <http://www.bios.net>) licence²⁵. These open-source-modelled licences are characterized by having no commercial restrictions other than covenants for sharing of improvements, relevant safety information and regulatory data and for preserving the opportunity for others to freely improve and use the technology.

Note added in proof: Analysis of seed progeny of all three plant species transformed with *S. meliloti* shows stable inheritance of the transgenic GUS and hygromycin resistance phenotypes, with approximately mendelian segregation. □

Methods

Bacteria

Sinorhizobium meliloti 1021 and *Rhizobium* sp. NGR234 (streptomycin-resistant ANU240) were provided by the Australian National University Collection by M. Djordjevic; *Mesorhizobium loti* MAFF303099 was obtained from C. Ronson, and a single gentamicin phosphotransferase gene was inserted (this paper); *A. tumefaciens* EHA105 was obtained from P. Maliga. Bacteria were grown on YM plates or TY liquid with streptomycin (200 µg ml⁻¹; *S. meliloti*, *Rhizobium* sp.) or gentamicin (20 µg ml⁻¹; *M. loti*) except for *A. tumefaciens*, which was grown on LB.

The genotype of the bacterial strains used was confirmed by PCR amplification using selective primer sets for the chromosomes and plasmids. Details of genotyping primers used in Fig. 2 are listed in Supplementary Table S1; sequences of other primers that have been used in our work are available upon request.

Plasmid construction

For the creation of modified Ti plasmids, suicide vectors were constructed by T/A cloning of a PCR-amplified fragment from the Ti plasmid of EHA105 into TOPO vector pCR2.1 (Invitrogen), corresponding to 1,316 bp encompassing the whole *virG* gene (pTiWB1) or a 995 bp part of the non-essential gene *moaA* (pTiWB3). The *oriT* fragment of plasmid RK2/RP4 was cloned in these vectors as an *Xba*I fragment following amplification from pSUP202 using primers EVS49 and EVS50²⁶. Electroporation of the suicide vectors into EHA105 resulted in integration of the whole vector by a single crossing-over event, thereby creating two functional *virG* genes (pTiWB1) or insertion of a second truncated *moaA* gene (pTiWB3). Co-integration was confirmed by PCR across the integration site and by Southern blotting showing duplication of the target locus.

The binary vector pCambia1405.1 was derived from pCambia1305.1 (GenBank accession number AF354045) by insertion of a spectinomycin/streptomycin marker into the *Sac*II site. To make pCambia1105.1, the kanamycin marker in pCambia1405.1 was removed and only the spectinomycin/streptomycin marker was retained. To create pCambia1105.1R, the *Pvu*II-*Pvu*II multiple cloning site (MCS) fragment in pCambia1105.1 was replaced by the larger *Pvu*II-*Pvu*II MCS from pCR2 after re-ligation of the *Eco*RI sites. The T-DNA region in these vectors was completely sequenced. Primers used to distinguish between pCambia1105.1 ('non-R') and pCambia1105.1R ('R') were 5'-CTGGCAGCAGGTTTC-3' and 5'-TACGGCGAGTTCGTAGGT-3', encompassing the MCS region.

Bacterial strain construction

Modified Ti plasmids were mobilized to *Sinorhizobium* and *Rhizobium* strains in a triparental mating with EHA105 containing either pTiWB1 or pTiWB3, an *E. coli* helper strain containing RP4-4, and the respective non-*Agrobacterium* acceptor strain. The pTiWB1 plasmid was mobilized to *M. loti* by mating with *Rhizobium* sp. NGR234 containing pTiWB1 and RP4-4. Colonies were then selected from transconjugants that did not contain RP4-4, as we discovered that its presence impaired plant transformation (data not shown). The binary vector pCambia1105.1R was subsequently introduced by electroporation. Either pCambia1105.1 or pCambia1405.1 was electroporated into *A. tumefaciens* strain EHA105 containing either the unmodified pEHA105 Ti plasmid or pTiWB1 or pTiWB3.

Plant transformation

Bacteria were grown on plates, resuspended to an absorbance at 600 nm of 1.0–1.5 and used for transformation of tobacco leaf discs (*Nicotiana tabacum* L. cv. Wisconsin 38) following the protocol in ref. 27, except using RMOP media solidified with 2.5 g l⁻¹ Phytigel and a co-cultivation time of 3, 5 or 5–11 days (for *A. tumefaciens*, *S. meliloti* and *M. loti*, and *Rhizobium* sp., respectively). The bacterial cultures used for explant treatment were analysed by PCR with *Agrobacterium*-specific primer sets, and by plating onto LB plates favouring the growth of *Agrobacterium* in comparison with other rhizobia (a more sensitive assay). In 'addback' experiments (in which a defined number of *A. tumefaciens* cells were deliberately added to rhizobial cultures), this allowed detection of one *A. tumefaciens* cell in 10⁹ rhizobial cells. This is several orders of magnitude more sensitive

than the PCR detection method. After co-culture, leaf discs were assayed for GUS activity (using 0.5 mg ml^{-1} X-glucA) or transferred to regeneration media containing antibiotics.

Rice (*Oryza sativa* L. cv. Millin) was transformed using seed-derived callus based on the protocol in ref. 28, with modifications (see Supplementary Information).

Arabidopsis thaliana L. cv. C24 was transformed using a floral-dip protocol¹⁵ with bacteria resuspended in infiltration medium containing Murashige and Skoog Basal Medium, 5.0% (w/v) sucrose, 0.05% (v/v) Silwet L-77 and 0.1% (w/v) MES, pH 5.7. The modified media contained Murashige and Skoog Basal Medium, 1.0% (w/v) sucrose, 0.02% (v/v) Silwet L-77 and 0.1% (w/v) MES, pH 7.0.

Analysis of transformed plants

Genomic DNA was extracted from the regenerated plants using DNAzol (Invitrogen), digested with *EcoRI* and used for Southern blotting with a ³²P-labelled *GUS*Plus probe. Plant DNA sequences flanking the T-DNA insertion site(s) were determined by restriction digest/adaptor ligation²⁹.

Received 26 November; accepted 29 December 2004; doi:10.1038/nature03309.

1. Gelvin, S. B. *Agrobacterium*-mediated plant transformation: the biology behind the "gene-jockeying" tool. *Microbiol. Mol. Biol. Rev.* **67**, 16–37 (2003).
2. Roa-Rodriguez, C. & Nottenburg, C. *Agrobacterium*-mediated transformation of plants. *CAMBIA technology landscape paper* (<http://www.bios.net/Agrobacterium>) (2003).
3. Van Montagu, M. & Jeff Schell (1935–2003): steering *Agrobacterium*-mediated plant gene engineering. *Trends Plant Sci.* **8**, 353–354 (2003).
4. Wood, D. W. *et al.* The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* **294**, 2317–2323 (2001).
5. Goodner, B. *et al.* Genome sequencing of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* **294**, 2323–2328 (2001).
6. Klein, D. T. & Klein, R. M. Transmittance of tumor-inducing ability to avirulent crown-gall and related bacteria. *J. Bacteriol.* **66**, 220–228 (1953).
7. Hooykaas, P. J. J., Klapwijk, P. M., Nuti, M. P., Schilperoort, R. A. & Rorsch, A. Transfer of the *Agrobacterium tumefaciens* Ti plasmid to avirulent *Agrobacteria* and to *Rhizobium ex planta*. *J. Gen. Microbiol.* **98**, 477–484 (1977).
8. van Veen, R. J. M., den Dulk-Ras, H., Bisseling, T., Schilperoort, R. A. & Hooykaas, P. J. J. Crown gall tumor and root nodule formation by the bacterium *Phyllobacterium myrsinacearum* after the introduction of an *Agrobacterium* Ti plasmid or a *Rhizobium* Sym plasmid. *Mol. Plant Microbe Interact.* **1**, 231–234 (1988).
9. van Veen, R. J. M., den Dulk-Ras, H., Schilperoort, R. A. & Hooykaas, P. J. J. Ti plasmid containing *Rhizobium meliloti* are non-tumorigenic on plants, despite proper virulence gene induction and T-strand formation. *Arch. Microbiol.* **153**, 85–89 (1989).
10. Goethals, K., Verecke, D., Jaziri, M., Van Montagu, M. & Holsters, M. Leafy gall formation by *Rhodococcus fascians*. *Annu. Rev. Phytopathol.* **39**, 27–52 (2001).
11. Lambert, B. *et al.* Identification and plant interaction of a *Phyllobacterium* sp., a predominant rhizobium of young sugar beet plants. *Appl. Environ. Microbiol.* **56**, 1093–1102 (1990).
12. Young, J. M., Kuykendall, L. D., Martinez-Romero, E., Kerr, A. & Sawada, H. Classification and nomenclature of *Agrobacterium* and *Rhizobium*—a reply to Farrand *et al.* (2003). *Int. J. Syst. Evol. Microbiol.* **53**, 1689–1695 (2003).
13. Galibert, F. *et al.* The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293**, 668–672 (2001).
14. Hood, E. E., Gelvin, S. B., Melchers, L. S. & Hoekema, A. New *Agrobacterium* helper plasmids for gene transfer to plants. *Transgenic Res.* **2**, 208–218 (1993).
15. Pueppke, S. G. & Broughton, W. J. *Rhizobium* sp. strain NGR234 and *R. fredii* USDA257 share exceptionally broad, nested host ranges. *Mol. Plant Microbe Interact.* **12**, 293–318 (1999).
16. Jefferson R. A., Harcourt R. L., Kilian A., Wilson, K. J. & Keese, P. K. Microbial β -glucuronidase genes, gene products and uses thereof. US patent 6,391,547 (2003).
17. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).
18. Watts, R. A. *et al.* A hemoglobin from plants homologous to truncated hemoglobins of microorganisms. *Proc. Natl Acad. Sci. USA* **98**, 10119–10124 (2001).
19. Kondo, N., Nikoh, N., Ijichi, N., Shimada, M. & Fukatsu, T. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. *Proc. Natl Acad. Sci. USA* **99**, 14280–14285 (2002).
20. Brown, J. R. Ancient horizontal gene transfer. *Nature Rev. Genet.* **4**, 121–132 (2003).
21. Martin, W. *et al.* Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl Acad. Sci. USA* **99**, 12246–12251 (2003).
22. Suzuki, K., Yamashita, I. & Tanaka, N. Tobacco plants were transformed by *Agrobacterium rhizogenes* infection during their evolution. *Plant J.* **32**, 775–787 (2002).
23. Reimann, C. & Haas, D. in *Bacterial Conjugation* (ed. Clewell, D. B.) 137–188 (Plenum, New York, 1993).
24. Nair, G. R., Liu, Z. & Binns, A. N. Reexamining the role of the accessory plasmid pATC58 in the virulence of *Agrobacterium tumefaciens* strain C58. *Plant Physiol.* **133**, 989–999 (2003).
25. Dennis, C. Biologists launch 'open-source movement'. *Nature* **431**, 494 (2004).
26. Stabb, E. V. & Ruby, E. G. RP4-based plasmids for conjugation between *Escherichia coli* and members of the Vibrionaceae. *Methods Enzymol.* **358**, 413–426 (2002).
27. Svab, Z., Hajdukiewicz, P. & Maliga, P. in *Methods in Plant Molecular Biology* (eds Maliga, P., Klessig, D. F., Cashmore, A. R., Gruissem, W. & Varner, J. E.) 55–77 (Cold Spring Harbor Laboratory Press, New York, 1995).
28. Hiei, Y., Ohta, S., Komari, T. & Kumashiro, T. Efficient transformation of rice (*Oryza sativa* L.) mediated by *Agrobacterium* and sequence analysis of the boundaries of the T-DNA. *Plant J.* **6**, 271–282 (1994).
29. Cottage, A., Yang, A., Maunders, H., de Lacy, R. C. & Ramsay, N. A. Identification of DNA sequences flanking T-DNA insertions by PCR walking. *Plant Mol. Biol. Rep.* **19**, 321–327 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was carried out within the Molecular Technologies Group of CAMBIA. We thank M. Irwin, J. Ward and H. Kilborn for their assistance with the plant transformation work. We thank P. Oger for sharing unpublished sequences of the pTiBo542 Ti plasmid, R. Wagner for comparisons of insertion sites with unpublished tobacco sequences, and M. Connett-Porceddu, P. Wenz and S. Hughes for discussions on the manuscript. This work was supported by grants from the Rockefeller Foundation, Horticulture Australia and Rural Industries R&D Corporation (RIRDC).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.A.J. (r.jefferson@cambia.org).

A universal trend of amino acid gain and loss in protein evolution

I. King Jordan¹, Fyodor A. Kondrashov², Ivan A. Adzhubei³, Yuri I. Wolf¹, Eugene V. Koonin¹, Alexey S. Kondrashov¹ & Shamil Sunyaev³

¹National Center for Biotechnology Information, NIH, Bethesda, Maryland 20894, USA

²Section of Evolution and Ecology, University of California at Davis, Davis, California 95616, USA

³Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA

Amino acid composition of proteins varies substantially between taxa and, thus, can evolve. For example, proteins from organisms with (G+C)-rich (or (A+T)-rich) genomes contain more (or fewer) amino acids encoded by (G+C)-rich codons^{1–4}. However, no universal trends in ongoing changes of amino acid frequencies have been reported. We compared sets of orthologous proteins encoded by triplets of closely related genomes from 15 taxa representing all three domains of life (Bacteria, Archaea and Eukaryota), and used phylogenies to polarize amino acid substitutions. Cys, Met, His, Ser and Phe accrue in at least 14 taxa, whereas Pro, Ala, Glu and Gly are consistently lost. The same nine amino acids are currently accrued or lost in human proteins, as shown by analysis of non-synonymous single-nucleotide polymorphisms. All amino acids with declining frequencies are thought to be among the first incorporated into the genetic code; conversely, all amino acids with increasing frequencies, except Ser, were probably recruited late^{5–7}. Thus, expansion of initially under-represented amino acids, which began over 3,400 million years ago^{8,9}, apparently continues to this day.

It is routinely assumed that extant proteins are in detailed equilibrium and their evolution is a stationary and reversible process: reciprocal fluxes of amino acid substitutions are equal, amino acid frequencies are constant, and nothing would change if time were to flow backwards^{10–12}. Accordingly, symmetrical substitution matrices are used for protein sequence alignment¹³.

We analysed 15 sets of three-way alignments of orthologous proteins encoded by triplets of closely related genomes (Table 1). At sites where the outgroup and one of the 'sister' genomes carry the same amino acid, whereas the other sister genome carries a different one, the amino acid ancestral for the sister genomes can be inferred. Accordingly, a specific, directed substitution can be assigned to the lineage of the sister that differs from the outgroup. For each set, we compared the 75 pairs of fluxes of forward and backward amino acid substitutions that can be caused by single-nucleotide replacements. The reciprocal fluxes differed by a factor of ≥ 1.5 for 575 of the 1,125 pairs (51%), and the differences were statistically significant.

cant ($P \leq 0.05$, with sequential rejective correction for multiple tests¹⁴) for 175 pairs (16%). For example, after the divergence of human and chimpanzee lineages, there were 102 Leu $\rightarrow$ Phe substitutions but only 57 Phe $\rightarrow$ Leu substitutions in the 7,552 com-

pared proteins. Thus, proteins are not in detailed equilibrium and their evolution is irreversible. Asymmetrical substitution matrices should be used for phylogeny-based alignment of multiple protein sequences¹⁵.

Table 1 Frequency changes of the strong gainer and loser amino acids in the 15 taxa

Taxon	Number of proteins	Amino acid distance*	Useful sites†	G+C content¶	Pu/Py imbalance#
Hominidae	7,552	0.4, 7.9, 7.9	8,549 (83)	-0.098	+0.051
Muridae	12,356	4.2, 11.3, 11.3	242,639 (73)	-0.003	+0.007
Saccharomyces	3,462	6.3, 10.4, 9.3	90,496 (79)	+0.135	+0.011
Pyrococcus	1,283	11.7, 15.7, 15.0	35,300 (75)	+0.160	-0.138
Escherichia	3,337	0.7, 1.2, 1.3	7,511 (95)	-0.188	-0.005
Salmonella	2,623	0.6, 8.8, 8.9	4,560 (75)	-0.162	-0.052
Buchnera	386	14.1, 24.4, 23.0	10,794 (60)	+0.504	-0.097
Vibrio	755	0.5, 13.5, 13.5	1,063 (36)	+0.057	-0.031
Pseudomonas	2,507	15.0, 19.1, 17.4	95,260 (69)	-0.508	+0.046
Bordetella	2,815	0.3, 0.7, 0.8	2,822 (99)	-0.453	+0.016
Helicobacter	768	1.7, 23.7, 23.5	2,333 (54)	+0.356	-0.073
Chlamydia	750	7.1, 20.8, 20.1	13,033 (61)	+0.272	+0.021
Bacillus	3,728	3.0, 4.8, 4.7	30,874 (87)	+0.160	-0.006
Streptococcus	1,065	0.6, 15.9, 15.9	1,589 (68)	+0.312	-0.015
Staphylococcus	1,750	0.4, 15.3, 15.3	1,754 (69)	+0.385	-0.049

	Substitutions to and from an amino acid‡								
	Cys	Met	His	Ser	Phe	Gly	Glu	Ala	Pro
Hominidae	137/72, +0.31	324/204, +0.23	297/206, +0.18	633/586, +0.04	231/115, +0.34	294/342, -0.08	232/386, -0.25	517/606, -0.08\$	204/437, -0.36
Muridae	2,547/1,729, +0.19	5,920/4,205, +0.17	5,702/4,752, +0.09	1,9904/1,8108, +0.05	4,461/3,895, +0.07	8,238/8,677, -0.03\$	8,056/11,269, -0.17	1,559/1,733, -0.05	7,350/9,883, -0.15
Saccharomyces	778/430, +0.29	1,768/1,455, +0.10	1,715/1,606, +0.03	7,776/7,567, +0.01	1,824/1,546, +0.08	3,370/2,511, +0.15	4,006/4,554, -0.06	5,336/4,950, +0.04	2,188/2,290, -0.02
Pyrococcus	35/21, +0.25	772/426, +0.29	277/244, +0.06	1,756/1,222, +0.18	680/598, +0.06	727/757, -0.02	1,663/2,836, -0.26	1,241/1,598, -0.13	211/416, -0.33
Escherichia	143/57, +0.43	238/155, +0.21	284/194, +0.19	738/549, +0.15	151/122, +0.11	239/334, -0.17	360/476, -0.14	531/1,129, -0.36	138/294, -0.36
Salmonella	67/11, +0.79	127/81, +0.22\$	122/74, +0.25\$	365/216, +0.26	84/57, +0.19	126/205, -0.24	176/202, -0.07	285/519, -0.29	53/167, -0.52
Buchnera	75/35, +0.36	166/102, +0.25	191/105, +0.29	654/587, +0.05	249/164, +0.21	99/149, -0.20\$	327/263, +0.11\$	403/396, +0.01	65/124, -0.31
Vibrio	4/0, +1.00	22/15, +0.19	16/10, +0.23	64/52, +0.10	18/17, +0.03	23/24, -0.02	45/70, -0.22	45/93, -0.35	10/21, -0.36
Pseudomonas	501/339, +0.19	2,985/1,090, +0.47	1,737/1,252, +0.16	6,289/4,277, +0.19	1,956/1,623, +0.09	2,759/3,186, -0.07	3,780/5,448, -0.18	6,623/9,261, -0.17	1,054/2,287, -0.37
Bordetella	89/14, +0.73	104/66, +0.22\$	113/75, +0.20\$	301/153, +0.33	50/51, -0.01	157/247, -0.22	102/125, -0.10	248/575, -0.40	83/188, -0.39
Helicobacter	15/11, +0.15	48/20, +0.41\$	55/24, +0.39\$	93/75, +0.11	31/28, +0.05	45/42, +0.03	44/67, -0.21	93/119, -0.12	14/27, -0.32
Chlamydia	107/59, +0.29	212/137, +0.22	249/150, +0.25	911/761, +0.09	238/160, +0.20	174/198, -0.07	307/512, -0.25	697/790, -0.06	116/224, -0.32
Bacillus	235/126, +0.30	960/722, +0.14	910/739, +0.10	2,044/1,861, +0.05\$	802/692, +0.07\$	898/932, -0.02	1,702/2,258, -0.14	1,911/2,117, -0.05\$	342/619, -0.29
Streptococcus	20/3, +0.74\$	27/26, +0.02	38/28, +0.15	114/61, +0.30	28/21, +0.14	28/50, -0.28	61/101, -0.25\$	71/128, -0.29	13/47, -0.57
Staphylococcus	10/1, +0.83	32/23, +0.16	30/25, +0.09	97/75, +0.13	44/31, +0.17	33/53, -0.23	81/93, -0.07	76/103, -0.15	6/47, -0.77
Average D, (standard, error of D)	+0.452 (0.07)	+0.219 (0.03)	+0.178 (0.03)	+0.135 (0.03)	+0.120 (0.02)	-0.098 (0.03)	-0.150 (0.03)	-0.163 (0.04)	-0.362 (0.04)

* Percentage of mismatches in alignments of orthologous proteins between sister genome 1 and sister genome 2, sister genome 1 and the outgroup, and sister genome 2 and the outgroup, respectively.
† The number of sites where the two sister proteins contain different amino acids, and the percentage (in parentheses) of such sites where the outgroup has the same amino acid as one of the sisters.
‡ The number of all substitutions creating (C) and removing (R), respectively, a particular amino acid are presented together with their normalized difference $D = (C - R)/(C + R)$. Out of 135 values of D, 89 (73 after the sequential rejective correction for multiple tests¹⁴) deviate from 0 significantly.
§ The values of D that deviated from 0 significantly ($P \leq 0.01$) are indicated.
|| The values of D that deviated from 0 significantly ($P \leq 0.05$), even after the multiple tests correction¹⁴, are indicated.
¶ The normalized difference between the number of substitutions creating and removing a G-C pair within a coding region.
The normalized difference between the number of substitutions creating and removing a purine within the coding strand. Pu, purine; Py, pyrimidine.

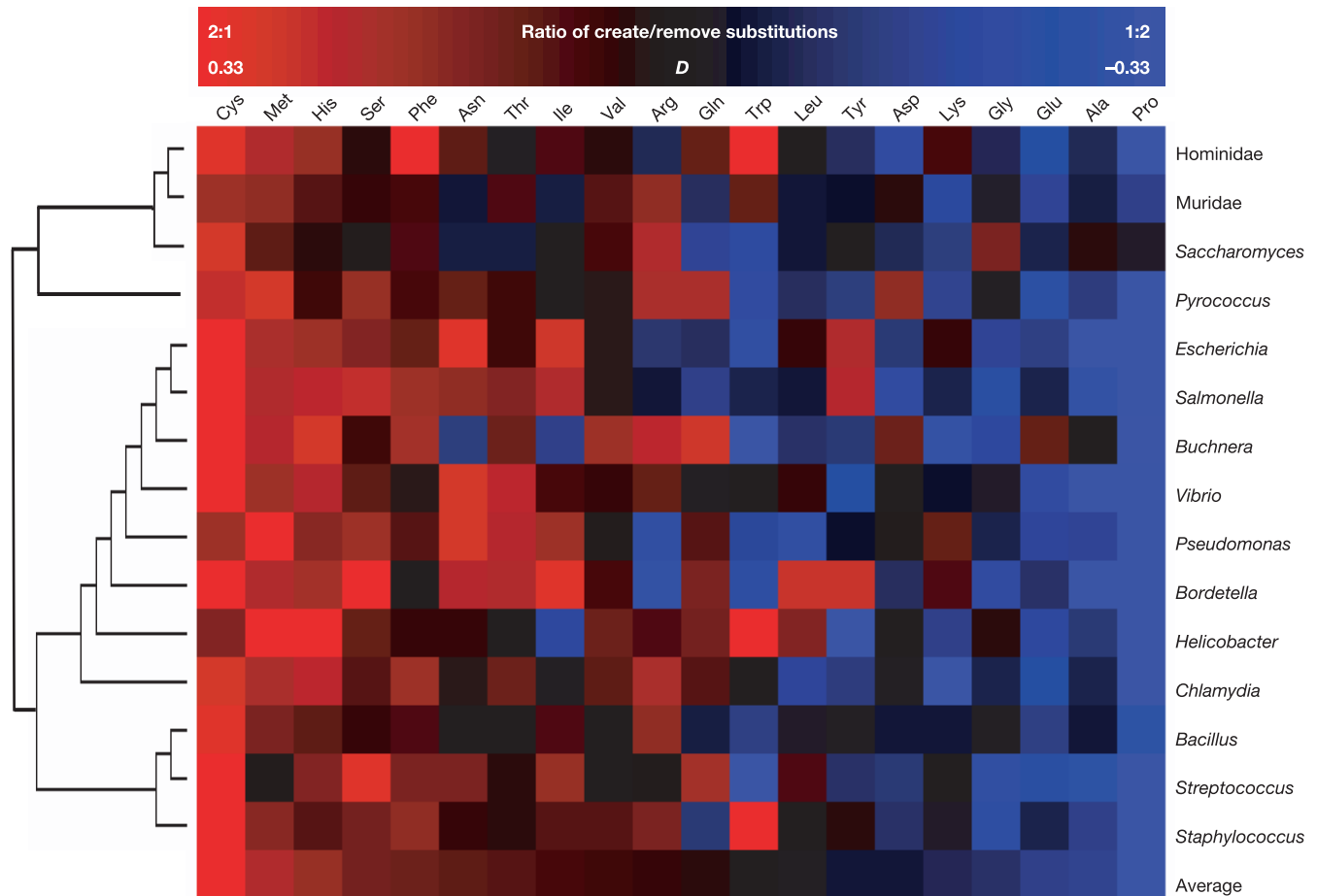


Figure 1 Normalized differences (D) between the number of substitutions creating (C) and removing (R) each amino acid ($D = (C - R)/(C + R)$) in 15 genome triplets.

$|D| = 0.33$ corresponds to a twofold difference between C and R . A tentative phylogenetic tree of the analysed taxa is shown to the left of the matrix.

Table 2 Gain and loss of amino acids in human SNPs

Amino acid	Number of polymorphisms*	Number of substitutions†	Relationship between divergence and polymorphism‡
Cys	1,046/484, + 0.367	137/72, + 0.311	0.88
Met	1,219/811, + 0.201	324/204, + 0.227	1.06
His	1,616/835, + 0.319	297/206, + 0.181	0.75
Ser	2,589/2,321, + 0.055	633/586, + 0.039	0.97
Phe	1,046/572, + 0.293	231/115, + 0.335	1.10
Asn	1,199/940, + 0.121	340/277, + 0.102	0.96
Thr	2,109/1,839, + 0.068	504/524, - 0.019	0.84
Ile	1,660/1,245, + 0.143	525/450, + 0.077	0.88
Val	2,459/2,038, + 0.094	725/674, + 0.036	0.89
Arg	2,321/3,665, - 0.225	570/670, - 0.081	1.34
Gln	1,510/1,151, + 0.135	359/291, + 0.105	0.94
Trp	557/221, + 0.432	61/25, + 0.419	0.97
Leu	2,178/1,690, + 0.126	394/397, - 0.004	0.77
Tyr	609/440, + 0.161	96/114, - 0.086	0.61
Asp	927/1,300, - 0.167	197/308, - 0.220	0.90
Lys	1,370/1,032, + 0.141	336/292, + 0.070	0.87
Gly	1,374/1,954, - 0.174	294/342, - 0.075	1.22
Glu	926/1,729, - 0.302	232/386, - 0.249	1.12
Ala	1,505/2,740, - 0.291	517/606, - 0.079	1.55
Pro	1,118/2,331, - 0.352	204/437, - 0.363	0.97

*For each amino acid, listed as in Fig. 1 in the order of the decreasing D values, the number of human non-synonymous SNPs where the amino acid is the derived allele (p_+), the number of SNPs where the amino acid is the ancestral allele (p_-), and the normalized difference D between these numbers is presented (all the normalized differences deviate from 0 significantly, $P < 10^{-9}$, except for Ser).

†For each amino acid, the number of substitutions creating (s_+) and removing (s_-) it since the human-chimpanzee divergence is shown, together with the normalized difference.

‡The ratio $r = (s_+/p_+)/(s_-/p_-)$ was computed for each amino acid. For a gainer, $r > 1$ if a large fraction of the substitutions that created it were driven by positive selection; conversely, if many substitutions removing a loser were driven by positive selection, $r < 1$ for it²⁴. Clearly, the values of r are not higher for gainers than for losers, arguing against a major role of positive selection.

Moreover, amino acid composition of proteins is far even from a total equilibrium (Fig. 1 and Table 1). With all substitutions taken into account, regardless of the number of nucleotide replacements required, five strong 'gainers' (Cys, Met, His, Ser and Phe) accrue in 14 or all 15 taxa, and frequencies of four strong 'losers' (Pro, Ala, Glu and Gly) decline in at least 13 taxa. Four amino acids are weak gainers, which either accrue in 11 taxa (Asn, Thr and Ile) or accrue slowly in all taxa (Val); in contrast, Lys is a weak loser, lost in ten taxa (Table 1). Frequencies of the remaining six amino acids (Arg, Gln, Trp, Leu, Tyr and Asp) evolve more erratically, although many individual changes are highly significant, reflecting taxon-specific mutational and selective pressures¹⁻⁴ (Supplementary Table 1).

The only conceivable source of systematic error in this analysis could be failure to identify the ancestral amino acid at a site where multiple substitutions occurred. This could disproportionately reduce the number of detected substitutions that remove evolvable and/or rare amino acids, thus creating a false impression that the frequencies of such amino acids increase¹⁶. However, a conservative (for our purposes) correction for multiple substitutions did not change the results substantially (Supplementary Methods and Supplementary Table 1). Also, the pattern of amino acid gain and loss in three taxa where the outgroup was close enough to the sister genomes to make misidentification of the ancestral state unlikely (*Escherichia*, *Bordetella* and *Bacillus*) was the same as in the remaining 12 taxa (Fig. 1 and Table 1). Furthermore, limiting the analysis to highly conserved regions of proteins did not affect this

pattern (data not shown). Finally, we examined human single-nucleotide polymorphisms (SNPs) polarized unambiguously by orthologous chimpanzee sequences. Because a majority of even non-synonymous human SNPs are effectively neutral¹⁷, they are expected to reveal a pattern of ongoing amino acid gain and loss in the human proteins similar to that during the course of human–chimpanzee divergence. Indeed, this was the case (Table 2), which corroborates our inferences from interspecies comparisons.

Because the sister species in all the 15 taxa are closely related, the observed pattern directly reflects only the last ~10 million years of evolution. We cannot conceive of a global external factor that could cause, during this time, parallel evolution of amino acid compositions of proteins in 15 diverse taxa that represent all three domains of life and span a wide range of lifestyles and environments. Thus, currently, the most plausible hypothesis is that we are observing a universal, intrinsic trend that emerged before the last universal common ancestor of all extant organisms. A comparison of extant proteins and reconstructed ancient proteins^{8,9} implied a long-term trend consistent with the trend observed in recent evolution (Supplementary Table 2), despite several substantial limitations of the deep reconstructions.

This trend in protein evolution is not driven by any simple trend at the DNA level. Among the losers, Pro, Ala and Gly are encoded by (G+C)-rich codons, so declining G+C content might, potentially, drive their loss^{1–4}. However, Arg and Trp, also encoded by (G+C)-rich codons, are not consistently lost, whereas Tyr and Lys, encoded by (A+T)-rich codons, are not gainers. Furthermore, frequencies of strong gainers or losers encoded by two codons, one containing two Gs or Cs and the other containing one G or C (Cys, His and Glu), cannot be directly affected by G+C content. Most importantly, there was no consistent pattern in evolution of the G+C content (which increases more often than it declines, in nine and six taxa, respectively; Table 1), purine/pyrimidine strand imbalance (Table 1), or several other DNA-level parameters (data not shown), in marked contrast to the uniformity observed at the protein level.

Proteins currently gain, on average, ~0.007 Cys and lose ~0.014 Pro residues per substitution (Table 3). Using a simple linear model

to extrapolate the observed dynamics backwards, we estimate that, since the time when the nine gainers first appeared, about two substitutions have been accepted per protein site, and proteins at that time mostly consisted of today's losers (Table 3). This epoch probably antedates the last universal common ancestor, which probably lived ~3.5 billion years ago¹⁸. Amino acid substitutions in murid rodents¹⁹, insects²⁰, vertebrates²⁰, diatoms²¹, Enterobacteriaceae²² and *Buchnera*²³ occur with average rates of 1.5, 1.5, 1, 2, 0.2 and 1.5 substitutions per site per billion years, respectively. Thus, the average number of substitutions per site since the last universal common ancestor is probably between 0.7 (the number suggested by Enterobacteriaceae; the proteins of the endosymbiont *Buchnera* evolve unusually quickly for bacteria²³) and 7, a range that includes the above estimate.

A forward extrapolation shows that the asymptotic frequencies of the gainers (losers), to be reached in the distant future, are much higher (lower) than their current frequencies. In extant proteins, gainers tend to be under-represented (except for Met), and losers are over-represented (except for Pro), relative to the frequencies corresponding to equal usage of all codons, but these differences are predicted to decline (Table 3).

Gainers (or losers) do not seem to be unusually similar to each other in terms of their physico-chemical properties, metabolic pathways, or codon assignments. However, a loser-rich and gainer-poor amino acid composition of primordial proteins may be explained historically by the order in which amino acids were recruited into the genetic code^{5–8}. All four strong losers (but only one of the five strong gainers, Ser) are among the ten amino acids produced in spark experiments imitating conditions for pre-biotic synthesis⁶. All strong losers, but none of the strong gainers, are among the eight amino acids found in the Murchison meteorite and thought to be abiogenic⁷. A consensus chronology of incorporation of amino acids into the genetic code, based on 60 criteria⁵, suggests that all strong losers are among the six most ancient amino acids, whereas four of the five strong gainers (except for Ser) are among the seven recruited last (Table 3; see also Supplementary Table 3). Notable exceptions are Tyr and Trp, assumed to have been recruited among the last but not accumulating consistently. Perhaps these

Table 3 Inferred evolutionary trajectories of amino acid frequencies

Amino acid*	Rate of gain/loss†	Current frequency	Asymptotic frequency‡	Frequency with unbiased usage of codons§	Number of substitutions since origin	Initial frequency¶	Consensus order ⁵ of recruitment into the genetic code
Cys	+0.0067 ± 0.00171	0.012	0.032	0.033	1.363	—	16
Met	+0.0088 ± 0.00126	0.025	0.038	0.016	1.595	—	19
His	+0.0073 ± 0.00107	0.022	0.032	0.033	1.592	—	14
Ser	+0.0167 ± 0.00356	0.062	0.080	0.098	1.597	—	6
Phe	+0.0042 ± 0.00091	0.042	0.053	0.033	4.067	—	17
Asn	+0.0073 ± 0.00268	0.040	0.048	0.033	1.985	—	13
Thr	+0.0091 ± 0.00282	0.051	0.061	0.066	1.957	—	9
Ile	+0.0089 ± 0.00500	0.068	0.077	0.049	2.120	—	12
Val	+0.0098 ± 0.00182	0.069	0.078	0.066	1.927	—	4
Arg	+0.0038 ± 0.00433	0.051	0.056	0.098	3.195	—	10
Gln	+0.0020 ± 0.00184	0.038	0.041	0.033	3.821	—	11
Trp	+0.0002 ± 0.00041	0.011	0.012	0.016	—	—	20
Leu	−0.0017 ± 0.00346	0.103	0.100	0.098	—	0.109	8
Tyr	−0.0005 ± 0.00118	0.032	0.031	0.033	—	0.034	18
Asp	−0.0039 ± 0.00229	0.052	0.047	0.033	—	0.074	3
Lys	−0.0065 ± 0.00333	0.059	0.051	0.033	—	0.092	15
Gly	−0.0063 ± 0.00253	0.072	0.059	0.066	—	0.093	1
Glu	−0.0137 ± 0.00244	0.065	0.048	0.033	—	0.136	7
Ala	−0.0239 ± 0.00862	0.081	0.058	0.066	—	0.242	2
Pro	−0.0139 ± 0.00262	0.043	0.022	0.066	—	0.101	5

* Assuming that u and v , the per site rates of substitution from and to an amino acid, respectively, are constant, x , the frequency of the amino acid, obeys the equation $dx/dt = v(1-x) - ux$. Thus, $x(t) = (x_0 - v/(u+v))\exp(-(u+v)t) + v/(u+v)$, with time measured in the units of the number of amino acid substitutions per site. We used values of u and v that are their arithmetic means across the 15 taxa. Obviously, forward and backward extrapolations of x obtained in this way should be treated with caution.

† The average number of gained or lost residues per one amino acid substitution per protein site for the 15 taxa (mean ± standard error; $n = 15$).

‡ In the distant future, the frequency of an amino acid will approach $v/(u+v)$.

§ Frequency of an amino acid attained if all non-stop codons were used equally; that is, $n/61$, where n is the number of codons that encode the amino acid.

|| The estimated average number of substitutions per site since the moment when the frequency of a gainer was 0.

¶ The frequency of a loser at the moment in the past since which two substitutions per average site occurred.

amino acids reached equilibrium concentrations more quickly than other late recruitments.

The initial frequency of a late-coming amino acid must be low because too many simultaneous changes in all proteins of the cell must be lethal. In particular, four of the five strong gainers (except for Ser) are thought to have been recruited by capturing codons that previously encoded other amino acids⁵. Such recruitments would need time to gain a foothold, at the expense of the primordial amino acids. Indeed, reconstructed ancient proteins were poor in Cys residues and some other late-coming amino acids, suggesting that they still accrued for some time after the last universal common ancestor (refs 8, 9 and Supplementary Table 2). However, it is remarkable that accumulation of late-coming amino acids continues to this day throughout the entire range of cellular life, affecting, among other things, the human–chimpanzee divergence and the genetic variability within human populations.

Accumulation of the gainers could not be driven solely by random drift because selectively neutral equilibrium would have been reached a long time ago. If gainers were favoured by positive selection, the rate of fixation of amino acid substitutions would be increased but the level of polymorphism would not. However, an analysis similar to the McDonald–Kreitman test²⁴ did not reveal any excess, in the course of human–chimpanzee divergence, in the rate of substitutions that add gainers or remove losers, relative to the density of corresponding human SNPs (Table 2). Probably, accumulation of the gainers was caused mostly by relaxation of negative selection if, for example, an ever-increasing fraction of sites where only Gly conferred high fitness initially became tolerant to Cys. Opening of extra sites to the gainers, if caused by epistatic interactions with other evolving sites^{25,26}, can be very slow. □

Methods

All 12 available triplets of complete prokaryotic genomes—for which the divergence of the amino acid sequences of orthologous proteins from the sister genomes (the closest pair) was within 0.3–15%, and the outgroup–sisters divergence was <25%—were extracted from the NCBI genome database using the Entrez retrieval engine (<http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html>). There are also three suitable triplets of eukaryotic genomes, resulting in the following 15 triplets, with sister genomes shown in parentheses. Hominidae: (*Homo sapiens*, *Pan troglodytes*), *Mus musculus*; Muridae: (*M. musculus*, *Rattus norvegicus*), *H. sapiens*; *Saccharomyces*: (*S. cerevisiae*, *S. paradoxus*), *S. mikatae*; *Pyrococcus*: (*P. horikoshii*, *P. abyssi*), *P. furiosus* DSM 3638; *Escherichia*: (*E. coli* K12, *E. coli* O157:H7), *E. coli* CFT073; *Salmonella*: (*S. typhimurium* LT2, *S. enterica enterica* serovar Typhi Ty2), *Shigella flexneri* 2a strain 301; *Buchnera*: (*B. aphidicola* strain APS, *B. aphidicola* strain Sg), *B. aphidicola* strain Bp; *Vibrio*: (*V. vulnificus* YJ016, *V. vulnificus* CMCP6), *V. parahaemolyticus* RIMD 2210633; *Pseudomonas*: (*P. syringae* pv. tomato strain DC3000, *P. putida* KT2440), *P. aeruginosa* PAO1; *Bordetella*: (*B. bronchiseptica* RB50, *B. parapertussis*), *B. pertussis* Tohama I; *Helicobacter*: (*H. pylori* 26695, *H. pylori* J99), *H. hepaticus* ATCC 51449; *Chlamydia*: (*C. trachomatis*, *C. muridarum*), *C. caviae* GPIC; *Bacillus*: (*B. anthracis* strain Ames, *B. cereus* ATCC 10987), *B. cereus* ATCC 14579; *Streptococcus*: (*S. pyogenes* M1 GAS, *S. pyogenes* MGA5315), *S. agalactiae* NEM316; *Staphylococcus*: (*S. aureus aureus* Mu50, *S. aureus aureus* MW2), *S. epidermidis* ATCC 12228.

Human–chimpanzee–mouse and *S. cerevisiae*–*S. paradoxus*–*S. mikatae* alignments were obtained from refs 27 and 28, respectively, and mouse–rat–human alignments were constructed as described in ref. 19. Triplets of orthologues, identified using the reciprocal best-hit approach²⁹, were aligned by ClustalW³⁰, except for pre-aligned primate and yeast proteins. To eliminate erroneous alignments and misidentified orthologues, triplets in which the divergence between two sister sequences significantly ($P < 0.001$) exceeded the minimal divergence between a sister and the outgroup sequence were discarded. To eliminate incorrectly aligned regions, which could originate from errors in genome assemblies or gene structure prediction, we analysed only sites embedded within a frame of length ten with at least six matches between the sister sequences and at least four matches between each of the sisters and the outgroup. All alignments are available at <ftp://ftp.ncbi.nih.gov/pub/koonin/Jordan/Cysteine>.

In each three-sequence alignment, only sites where the outgroup and one of the sisters carries the same amino acid but the two sisters carry different amino acids were analysed. Sites where the outgroup differed from both sisters were ignored because the ancestral state of the sisters cannot be inferred. The 15 20 × 20 matrices of amino acid substitutions are available at <ftp://ftp.ncbi.nih.gov/pub/koonin/Jordan/Cysteine>. Fluxes of reciprocal substitutions caused by single-nucleotide replacements were compared and the statistical significance of the observed deviations from detailed equilibrium was estimated for each of the 75 pairs of reciprocal fluxes using a proportion test, after which sequential rejective Bonferroni procedure³¹ for multiple tests was applied to the whole set of results. Owing to the close relatedness of the sister genomes, substitutions that require two or three nucleotide replacements comprised, on average, only 6.8% of all substitutions (from 0.7%

in Hominidae to 17% in *Pseudomonas*) and, thus, were too uncommon to compare individual fluxes. The total gain or loss of an amino acid was calculated by counting all substitutions to and from it (Table 1; see also Supplementary Table 1).

This approach relies on close similarity between the sister and the outgroup sequences. Otherwise, it becomes prone to systematic error stemming from two sources: (1) the probability of identical substitutions in the outgroup and one of the sister lineages or of multiple substitutions in a lineage becomes substantial, leading to erroneous inference of the ancestral state and, thus, to misidentification of the substitution; and (2) the fraction of sites where both sisters differ from the outgroup becomes substantial, and exclusion of such sites biases the count of substitutions. To deal with these effects, we used a statistical correction for multiple substitutions (Supplementary Methods). The correction was devised to be conservative with respect to the asymmetry of reciprocal substitutions; that is, it tends to underestimate gain or loss of amino acids.

Human non-synonymous SNPs were extracted from dbSNP build 120 (<http://ncbi.nlm.nih.gov/SNP/>). The ancestral amino acid for each SNP was inferred using the orthologous chimpanzee sequence from the UCSC chimpanzee genome assembly, version PanTro1 (<http://hgdownload.cse.ucsc.edu/goldenPath/panTro1/bigZips/>). The 29,262 SNP sites were analysed as described for substitutions between species.

Received 14 September; accepted 15 December 2004; doi:10.1038/nature03306.

Published online 19 January 2005.

- Sueoka, N. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc. Natl Acad. Sci. USA* **47**, 1141–1149 (1961).
- Gu, X., Hewett-Emmett, D. & Li, W.-H. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica* **103**, 383–391 (1998).
- Knight, R. D., Freeland, S. J. & Landweber, L. F. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **4**, Research0010.1 (2001).
- Wang, H.-C., Singer, G. A. C. & Hickey, D. A. Mutational bias affects protein evolution in flowering plants. *Mol. Biol. Evol.* **21**, 90–96 (2004).
- Trifonov, E. N. The triplet code from first principles. *J. Biomol. Struct. Dyn.* **22**, 1–11 (2004).
- Miller, S. L. Which organic compounds could have occurred on the prebiotic earth? *Cold Spring Harb. Symp. Quant. Biol.* **52**, 17–27 (1987).
- Cronin, J. R. & Pizzarello, S. Amino acids in meteorites. *Adv. Space Res.* **3**, 5–18 (1983).
- Brooks, D. J., Fresco, J. R., Lesk, A. M. & Singh, M. Evolution of amino acid frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code. *Mol. Biol. Evol.* **19**, 1645–1655 (2002).
- Brooks, D. J. & Fresco, J. R. Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. *Mol. Cell. Proteom.* **1**, 125–131 (2002).
- Muller, T. & Vingron, M. Modeling amino acid replacement. *J. Comp. Biol.* **7**, 761–776 (2000).
- Goldman, N. & Whelan, S. A novel use of equilibrium frequencies in models of sequence evolution. *Mol. Biol. Evol.* **19**, 1821–1831 (2002).
- Veerassamy, S., Smith, A. & Tillier, E. R. M. A transition probability model for amino acid substitutions from blocks. *J. Comp. Biol.* **10**, 997–1010 (2003).
- Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices. *Adv. Prot. Chem.* **54**, 73–97 (2000).
- Holm, S. A simple sequentially rejective multiple test procedure. *Stand. J. Stat.* **6**, 65–70 (1979).
- Feng, D. F. & Doolittle, R. F. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. *Methods Enzymol.* **266**, 368–382 (1996).
- Eyre-Walker, A. Problems with parsimony in sequences of biased base composition. *J. Mol. Evol.* **47**, 686–690 (1998).
- Sunyaev, S. et al. Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597 (2001).
- Tice, M. M. & Lowe, D. R. Photosynthetic microbial mats in the 3,416-Myr-old ocean. *Nature* **431**, 549–552 (2004).
- Rat Genome Sequencing Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–535 (2004).
- Zdobnov, E. M. et al. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**, 149–159 (2002).
- Sorhannus, U. & Fox, M. Synonymous and nonsynonymous substitution rates in diatoms: A comparison between chloroplast and nuclear genes. *J. Mol. Evol.* **48**, 209–212 (1999).
- Ochman, H. & Wilson, A. C. Evolution in bacteria—evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* **26**, 74–86 (1987).
- Clark, M. A., Moran, N. A. & Baumann, P. Sequence evolution in bacterial endosymbionts having extreme base compositions. *Mol. Biol. Evol.* **16**, 1586–1598 (1999).
- Smith, N. G. C. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024 (2002).
- Fitch, W. M. & Markowitz, E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4**, 579–593 (1970).
- Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA* **99**, 14878–14883 (2002).
- Clark, A. G. et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 233–234 (2003).
- Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements S.S. and I.A.A. were supported by the Genome Canada Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.S.K. (kondrashov@ncbi.nlm.nih.gov) or S.S. (ssunayev@rics.bwh.harvard.edu).

Contributions of an avian basal ganglia–forebrain circuit to real-time modulation of song

Mimi H. Kao, Allison J. Doupe & Michael S. Brainard

Keck Center for Integrative Neuroscience, Departments of Physiology and Psychiatry, University of California, San Francisco, California 94143-0444, USA

Cortical–basal ganglia circuits have a critical role in motor control and motor learning¹. In songbirds, the anterior forebrain pathway (AFP) is a basal ganglia–forebrain circuit required for song learning and adult vocal plasticity but not for production of learned song^{2–5}. Here, we investigate functional contributions of this circuit to the control of song, a complex, learned motor skill. We test the hypothesis that neural activity in the AFP of adult birds can direct moment-by-moment changes in the primary motor areas responsible for generating song. We show that song-triggered microstimulation in the output nucleus of the AFP induces acute and specific changes in learned parameters of song^{6,7}. Moreover, under both natural and experimental conditions, variability in the pattern of AFP activity is associated with variability in song structure. Finally, lesions of the output nucleus of the AFP prevent naturally occurring modulation of song variability. These findings demonstrate a previously unappreciated capacity of the AFP to direct real-time changes in song. More generally, they suggest that frontal cortical and basal ganglia areas may contribute to motor learning by biasing motor output towards desired targets or by introducing stochastic variability required for reinforcement learning.

Song is a complex, learned motor skill that involves the precise coordination of vocal and respiratory musculature in order to produce highly stereotyped renditions of a memorized song model. Two pathways have been identified in the songbird forebrain that contribute to this behaviour (Fig. 1a): (1) a motor pathway that is required throughout life for normal song production⁸; and (2) a basal ganglia–forebrain circuit⁹ (the AFP) that is necessary for song learning and plasticity but not for the production of adult song^{2–5}. The AFP converges with the motor pathway at the premotor nucleus RA (robust nucleus of the arcopallium). Although lesion studies have demonstrated that the AFP is required for vocal motor plasticity^{2–5}, its function in motor control and learning remains unclear.

Previous modelling work has hypothesized that a critical contribution of the AFP to song plasticity relies on its capacity to modulate ongoing song^{10,11}. Such a modulatory influence might serve to introduce into song the stochastic variability that is necessary for reinforcement learning¹¹, or to systematically bias song towards desired targets in a manner that eventually becomes encoded in the song motor pathway¹⁰. Consistent with this hypothesis, the AFP exhibits patterned neural activity in singing birds^{12,13}.

However, previous studies have failed to demonstrate any direct contribution of the AFP to the production of adult birdsong. In adult zebra finches, lesions of LMAN (lateral magnocellular nucleus of the anterior nidopallium), the output nucleus of the AFP (Fig. 1a), have not been reported to affect ongoing song. Moreover, whereas stimulation in the motor pathway of quiescent birds can elicit vocalizations^{14,15}, comparable microstimulation in LMAN does not. Finally, a previous study did not report any gross effects of LMAN microstimulation on song¹⁴. These findings indicate that LMAN is not an obligatory premotor structure for song, but they leave unresolved whether, and how, patterned activity from the AFP modulates song production.

Here, we examine the hypothesis that neural activity in the AFP contributes to vocal control by using microstimulation triggered by real-time song to manipulate activity in LMAN during precisely targeted parts of a song (Fig. 1b). We report that artificially altering the pattern of activity in LMAN induces acute changes in the structure of individual song elements, or ‘syllables’, without altering the order or structure of ensuing syllables. Such changes include increases and decreases in sound frequency (a learned parameter of song; ref. 6), as well as increases and decreases in sound amplitude (a parameter of song that is precisely controlled; refs 7, 16). Figure 1c illustrates a representative experiment. Every time the bird sang a rendition of a stereotyped sequence of syllables, or ‘motif’ (‘abcdef’), syllable ‘b’ was detected using a real-time template-matching algorithm. After detection, microstimulation was delivered in LMAN on randomly selected trials (in this case, during the first motif rendition, but not the second). In this experiment, LMAN stimulation caused a systematic downward shift in the fundamental frequency of syllable ‘c’. Figure 1d illustrates a second experiment using the same paradigm in which LMAN microstimulation caused a systematic decrease in the loudness of the targeted syllable. Such acute, stimulation-induced changes in syllable structure were observed for 18 of 20 sites in LMAN of five birds (Supplementary Tables 1 and 2).

Evoked changes in syllable structure were tightly locked to the delivery of stimuli. In cases where latency could be assessed, the mean latency between the onset of a stimulus and an effect on syllable structure was 50 ms, but it could be as short as 35 ms (see Supplementary Fig. 1). In addition, the effects of stimulation typically terminated within 60–70 ms after the end of the stimulus train (Fig. 1d). LMAN projects directly to the song motor nucleus RA, which is thought to provide motor commands that control the precise structure of individual song elements with a latency of 40–45 ms^{14,17}. Although the exact mechanisms and pathways underlying the influence of LMAN on song remain to be determined, both the rapidity with which LMAN stimulation could drive changes in syllable structure and the rapid termination of its effects are consistent with a direct modulation of RA by LMAN.

In a given experiment, stimulation in LMAN typically induced systematic changes in syllable structure, rather than a general degradation of song structure or an enhancement of song variability. When stimulation elicited a significant shift in the mean fundamental frequency of a syllable, it had no effect on the variability of the fundamental frequency in 59% of cases (note the standard deviation (s.d.) of the histograms in Fig. 2b), increased the variability in 30% of cases (Fig. 1c), and reduced the variability in the remaining 11% of cases. Thus, the predominant effect of artificially imposing a fixed pattern of activity in LMAN during singing was to shift systematically the mean value of syllable parameters, rather than to grossly disrupt song.

There was significant specificity to the changes elicited by microstimulation in LMAN. For 13 of 18 sites, a fixed pattern of stimulation had qualitatively different effects when delivered during different syllables (Supplementary Table 2). Figure 2a shows a case in which stimulation delivered at a fixed site in LMAN caused

an increase in the amplitude of one syllable, a decrease in the amplitude of a second syllable, and little change to a third syllable. Moreover, when we explicitly varied the site of stimulation in LMAN, while holding all other parameters constant, we could elicit qualitatively different effects on a given syllable. Figure 2b shows a case where stimulation at a dorsal site in LMAN caused a significant increase in the mean fundamental frequency, while stimulation at a more ventral site in LMAN caused a significant decrease in the fundamental frequency of the same syllable. These results suggest that LMAN may be functionally compartmentalized, consistent with known topographic projections from LMAN to RA¹⁸.

The observed changes in syllable structure were restricted to stimulation in LMAN. Stimulation that induced significant shifts in frequency when delivered in LMAN was ineffective when applied at control sites 400–1,000 μm dorsal to LMAN (Fig. 2c). At higher current intensities, stimulation outside of LMAN rarely caused significant changes in syllable structure (Fig. 2c). In contrast, we found that stimulation with higher current intensities both within LMAN and at control sites dorsal to LMAN could induce the suspension of ongoing motifs, as seen previously with stimulation of the premotor nucleus HVC¹⁴. Song suspensions occurred at a

median current intensity that was 233% of that required to elicit significant effects on syllable structure (range: 200–300%, $n = 8$). Because song suspensions could be elicited from control sites outside of LMAN, they are not likely to reflect the specific activation of RA, and may instead result from antidromic activation of nucleus HVC (Fig. 1a), which sends projections through the anterior forebrain^{14,19}.

The acute changes in syllable structure induced by artificially altering activity in LMAN indicate that neural activity in LMAN can modulate ongoing motor performance on a moment-by-moment basis. These results raise the question of whether the natural pattern of activity in LMAN modulates the motor pathway and subsequent song. To examine this possibility, we compared song produced by birds in two behavioural conditions in which the neural activity in the AFP is known to differ: when a male bird sings alone ('undirected' song), activity in LMAN is greater in magnitude and more variable in pattern across renditions than when it sings to another bird ('directed' song)^{12,20} (Fig. 3a–c, e). We found that the mean fundamental frequency of syllables did not differ systematically between these two conditions, but the variability in the fundamental frequency was significantly greater during undirected song (Fig. 3d, f). Moreover, across birds, there was a significant corre-

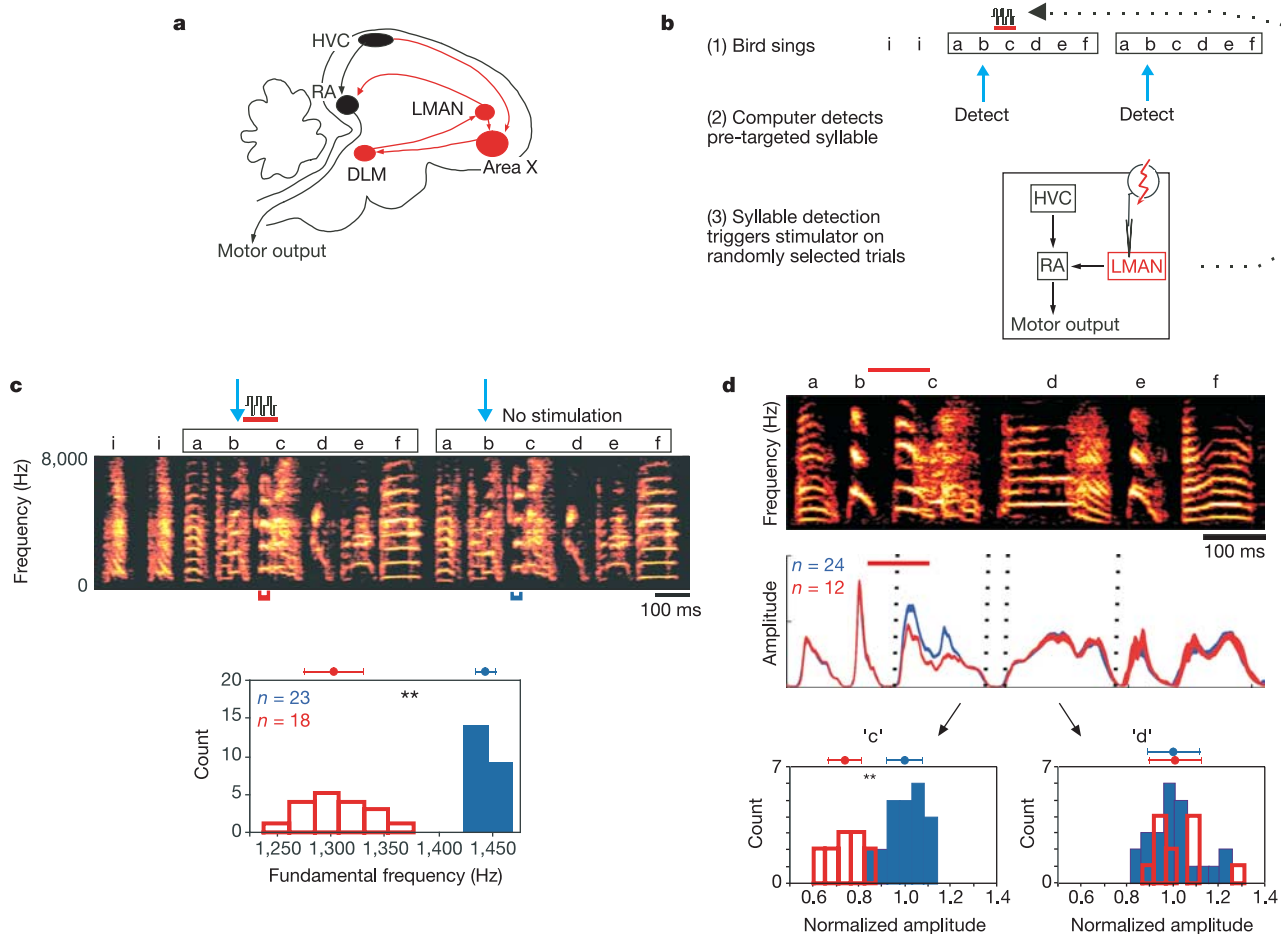


Figure 1 Song-triggered microstimulation in LMAN elicits acute changes in learned parameters of syllable structure. **a**, Song system. The motor pathway includes HVC and RA. The AFP consists of area X, the medial portion of the dorsolateral thalamus (DLM) and LMAN. **b**, Experimental design. 'i' indicates an introductory element that usually occurs at the start of a song bout before the motif(s) ('abcdef'). **c**, Stimulation-induced shift in frequency. Top: spectrogram illustrating two trials. LMAN was stimulated (30 μA ; red bar) during syllable 'c' on random trials. Bottom: fundamental frequency for syllable 'c' for

control (blue) and stimulation trials (red). Stimulation caused a 10% decrease in fundamental frequency (means (circles) $\pm$ s.d. (bars); $P < 0.0001$). **d**, Stimulation-induced change in amplitude. Top: spectrogram of another bird's song. Middle: mean amplitude waveforms ($\pm$ s.e.m.) for control (blue) and stimulation trials (red). Bottom: stimulation (40 μA) significantly reduced the amplitude of syllable 'c' ($\sim 27\%$; $P < 0.0001$), but not that of the next syllable, 'd' ($P = 0.74$).

lation between the magnitude of changes in LMAN variability and the magnitude of changes in song variability ($R^2 = 0.83$; Supplementary Fig. 2). The observed correspondence between variability in LMAN activity and variability in syllable structure (Fig. 3e, f) is consistent with a direct modulatory influence of LMAN on song under natural conditions.

To test explicitly whether trial-by-trial variability in LMAN activity can give rise to variability in motor output, we altered the pattern of LMAN activation across song renditions by varying the current intensity. Stimulation with a fixed current intensity tended to elicit a systematic change in fundamental frequency (for example, Figs 1c and 2b); however, varying the activity of LMAN neurons at a single site across motif renditions caused a significant increase in the variability of the fundamental frequency (Fig. 4a, b; Supplementary Table 3). Thus, the artificial introduction of variability in LMAN activity was sufficient to recapitulate one natural difference between directed and undirected song (compare Fig. 3f with Fig. 4b).

If activity from LMAN is indeed responsible for the difference in song variability between behavioural contexts, then removing this activity should eliminate the observed difference. We tested this by measuring the variability of syllable structure in directed and undirected songs of five birds before and after bilateral lesions of LMAN. Before lesions, for each bird, the variability in the fundamental frequency was significantly greater during undirected song (Fig. 4c; 2–5 syllables per bird). Immediately after the lesions, this context-dependent difference in song variability was eliminated (1–3 days after lesion; Fig. 4c). Thus, LMAN is necessary for modulating naturally occurring differences in song variability.

Our results suggest that a critical contribution of the AFP to song plasticity may derive from its previously unappreciated role in acutely modulating activity in the premotor nucleus RA. Neural

activity in LMAN could contribute to the adaptive modification of song in juvenile and adult birds in at least two ways⁴. One possibility is that LMAN provides an instructive signal to RA that systematically biases the motor pathway towards a particular goal^{3,4,10}. Consistent with this idea, we found that for a given syllable, a fixed pattern of LMAN activation typically induced specific changes in syllable structure. This specificity suggests that the AFP has the capacity to selectively bias independent components of song towards desired targets. An analogous instructive role has been postulated to be one function of descending inputs from frontal cortical and striatal regions in other vertebrate systems^{21,22}.

A second possibility is that neural activity in LMAN contributes to plasticity by generating variability in the motor pathway and subsequent song output^{3,11}. Consistent with this idea, we found that both naturally occurring and experimentally induced variability in LMAN activity is associated with variability in song output. Moreover, although LMAN is not required for the production of learned song, it is necessary for the state-dependent changes in the variability of syllable structure. In the context of feedback-based reinforcement learning, trial-by-trial variability in motor output is required in order for evaluation mechanisms to selectively reinforce the patterns of motor activity that produce the desired behaviour^{11,23}. Once a behaviour is learned, the ability to modify it remains important, both for feedback-based correction of errors that may result from changes in the periphery (for example, altered muscle tone or innervation) and for continued motor exploration to optimize the behaviour.

These two models for the contribution of LMAN to vocal plasticity are not mutually exclusive. Rather, the influence of the AFP will depend on its actual pattern of activity during singing. When AFP activity is stereotyped from one song rendition to the next (for example, during directed song), our results suggest that it

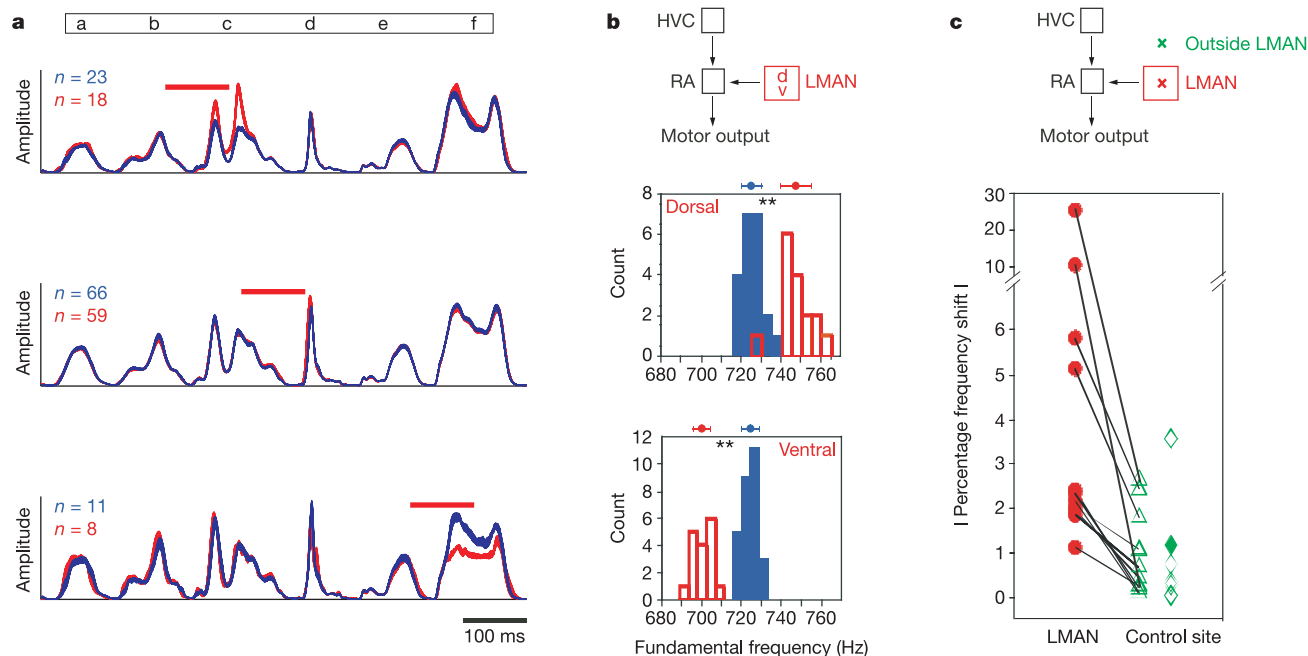


Figure 2 Specificity of stimulation. **a**, Stimulation at one site could elicit qualitatively different effects on different syllables. A fixed pattern of stimulation (30 μ A; red bar) increased the amplitude of syllable 'c', decreased the amplitude of syllable 'f', and had little effect on other syllables (for example, 'd'). Red, stimulation; blue, control. **b**, Stimulation at different sites in LMAN could induce qualitatively different effects for a given syllable. At one site, stimulation (20 μ A) increased the fundamental frequency (means (circles) $\pm$ s.d. (bars); $P < 0.0001$), whereas at a more ventral site, the same

stimulation decreased the fundamental frequency ($P < 0.0001$). **c**, Evoked changes were specific to stimulation in LMAN. Lines connect data points where the same stimulation was delivered inside (red circles) and outside (green triangles) of LMAN. Significant changes of fundamental frequency (filled symbols) were observed only inside LMAN. When current intensity was increased at control sites (green diamonds), significant effects were elicited in one of seven cases.

should systematically bias vocal production. In contrast, when AFP activity is variable across motif renditions (for example, during undirected song), it should contribute to variability in motor output, which may be a critical component of motor exploration during the process of vocal learning and/or maintenance. According to both of these models, the absence of patterned activity from the AFP accounts for the elimination of long-term vocal plasticity by lesions of LMAN^{4,5}.

The anatomical substrates for the influence of LMAN activity on the motor pathway are individual premotor neurons in RA that receive inputs from both HVC and LMAN²⁴. This convergence of inputs on the same RA neurons provides a cellular locus where LMAN activity can influence synaptic connections and/or transmission in RA²⁵. Moreover, LMAN input to RA is unusual in that it is mediated predominantly by NMDAR (*N*-methyl-D-aspartate receptors)^{24,26}, which are known to be important for structural and/or functional plasticity. NMDAR-mediated synaptic responses

evoked by LMAN terminals in RA are well situated for modulating the gain of active synapses in RA during singing. Modulation of ongoing motor activity has been observed in other systems^{27,28}, and may provide a general mechanism for enabling plasticity in motor circuits.

Several lines of evidence suggest that cortical–basal ganglia circuits contribute to the selection and sequencing of behaviour during the learning and performance of motor skills. Disorders of the basal ganglia can be characterized, in part, as either resulting in too much movement (for example, Huntington's disease) or in too little movement (for example, Parkinson's disease), suggesting that activity in cortical–basal ganglia circuits adjusts the gain of motor output. Moreover, altered patterns of activity in the basal ganglia (either experimentally induced or in disease states) can affect the degree of stereotypy versus variability in motor performance^{29,30}. We suggest that in songbirds, the AFP can modulate ongoing motor activity and drive specific changes in motor output on a moment-

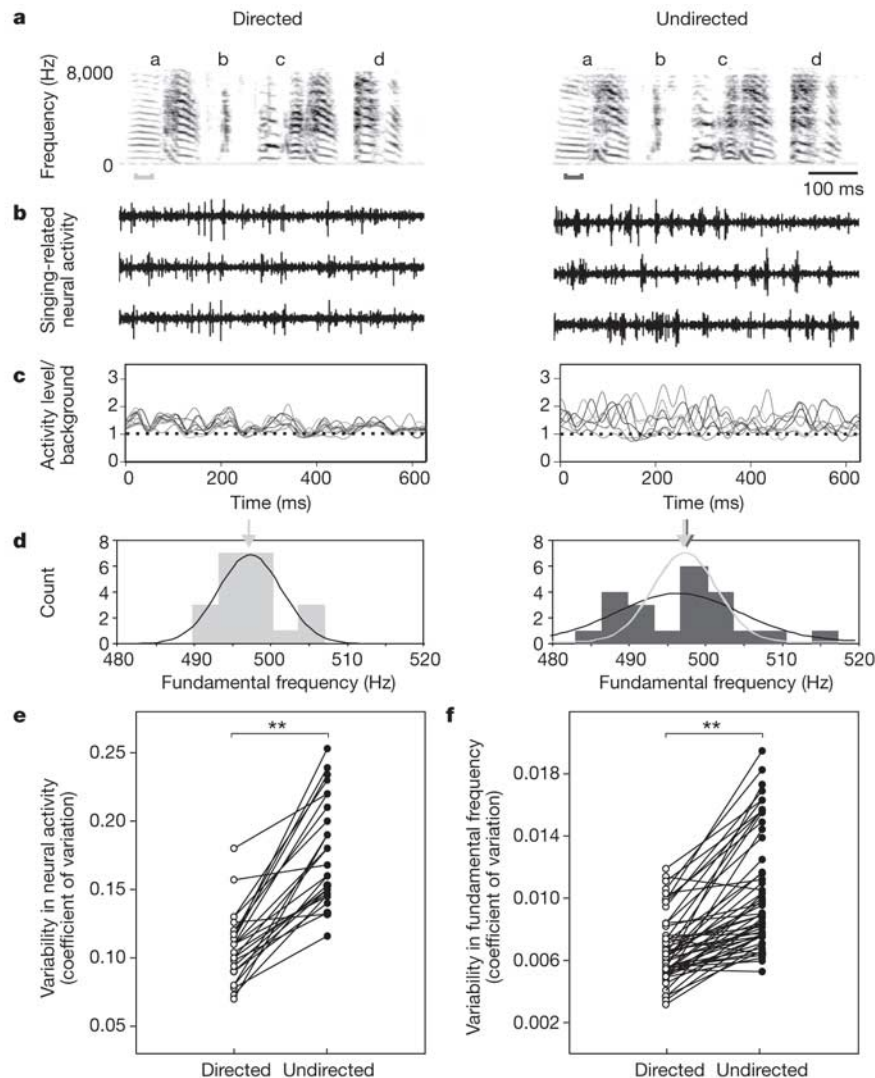


Figure 3 Context-dependent changes in variability. **a**, Spectrograms of the motif 'abcd' during directed (left) and undirected (right) singing. **b**, LMAN multi-unit activity during three renditions of the motif. **c**, LMAN activity waveforms for ten renditions of the motif. Note that the singing-related activity (normalized against background activity level) is greater than the background activity level; the dotted line represents the mean background activity level during non-singing periods. **d**, Histograms and gaussian fits of fundamental frequency for syllable 'a'. Mean fundamental frequency (arrows) did not

differ (497.4 and 496.7 Hz; $P = 0.54$), but variability in fundamental frequency was greater in undirected song (s.d., 4.17 and 8.10; $P = 0.004$). The gaussian fit for directed song is overlaid in grey on the gaussian fit for undirected song (right). **e**, Variability in activity was greater in undirected song ($n = 29$ sites, 11 birds; $P < 0.0001$). **f**, Variability in fundamental frequency was greater in undirected song (50 of 53 syllables, 18 birds; $P < 0.0001$).

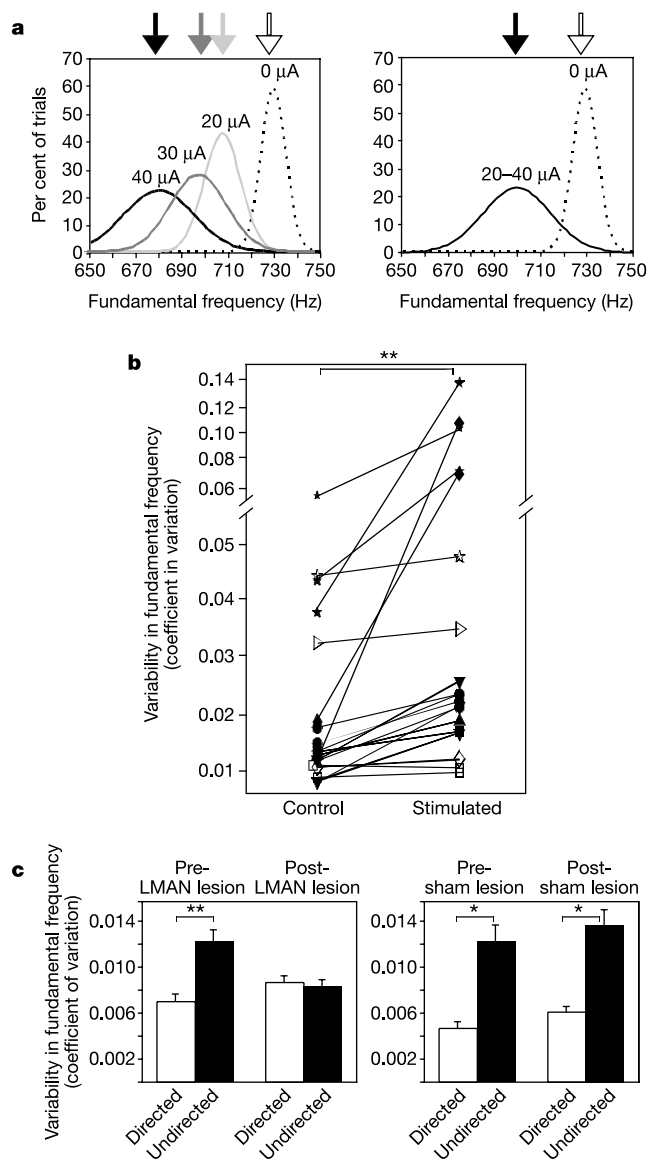


Figure 4 Contributions of the AFP to real-time song modulation. **a**, Left: gaussian fits of fundamental frequency for one syllable for control (dotted) and stimulated trials of different current intensities. Right: varying stimulation across trials increased the variability in fundamental frequency relative to controls ($P = 0.022$). **b**, Variability in fundamental frequency for control and stimulated trials in which different current intensities were delivered at the same site. Symbols denote syllables (1–3 per bird). Differential activation of LMAN increased the variability in the fundamental frequency in 18 of 19 cases ($P < 0.0001$). **c**, LMAN lesions eliminate context-dependent differences in song variability. Bars indicate the mean coefficient of variation of the fundamental frequency ($\pm$ s.e.m.) for syllables during directed and undirected song before and after LMAN lesions (19 syllables, five birds: pre, $P < 0.0001$; post, $P = 0.167$) or sham lesions (six syllables, two birds: pre, $P = 0.03$; post, $P = 0.03$).

by-moment basis, perhaps by adjusting the gain of active synapses in the motor pathway. In addition, we show that introduction of variability in the activity of LMAN (the output nucleus of the AFP) can drive variability in motor performance. Finally, we find that lesions of LMAN eliminate the naturally occurring context-dependent modulation of song variability. We suggest that a key contribution of frontal–basal ganglia circuits to motor learning and performance relies on the capacity of these circuits to bias motor output towards specific targets and to introduce variability in motor output.

Methods

Subjects

Juvenile (>80 days) and adult (>125 days) male zebra finches (*Taeniopygia guttata*) were used for experiments. Birds were selected on the basis of size, singing frequency and song complexity, and were isolated in a sound-attenuating chamber. All procedures were approved by the UCSF Institutional Animal Care and Use Committee.

Physiological recording

Surgical procedures were performed as described previously¹³. Briefly, a lightweight microdrive (UCSF and Caltech Machine Shops) carrying two metal electrodes (2–5 M Ω) was positioned stereotactically above LMAN in the right hemisphere. A reference electrode was implanted within 2 mm of LMAN, and the microdrive and a connector socket were secured to the bird's skull.

During experimental sessions, a flexible lead terminating in an operational amplifier was connected to the socket on the bird's head, and the other end was connected to a commutator¹³. Neural signals were amplified and filtered between 300 and 10,000 Hz. Acoustic signals were recorded with a small microphone above the birdcage and filtered between 200 and 9,000 Hz. Custom-written acquisition software (C. Malek and A. Leonardo, Caltech, and C. Roddey, UCSF) recorded the acoustic and neural signals before and after the sound amplitude crossed a threshold level, and the bird's behaviour was monitored by a video camera.

Electrodes were positioned either at control sites 400–1,000 μ m dorsal to LMAN ($n = 3$ birds) or at sites in LMAN selected on the basis of their characteristic singing-related activity¹³ ($n = 5$ birds). At sites at least 400 μ m dorsal to LMAN, there was no conspicuous change in multi-unit firing during singing when compared with the spontaneous neural activity during non-singing periods. At the conclusion of experiments, site locations were confirmed in 40- μ m Nissl-stained brain sections by their positions relative to the depth of marker lesions.

Song-triggered microstimulation

To deliver electrical stimuli reproducibly during specific parts of a song, custom-written software (J. Houde and C. Roddey, UCSF) compared the bird's vocalizations with pre-defined spectral templates for targeted song elements in real time. Detection triggered unilateral microstimulation via the same electrodes used for recording activity. For both fixed current amplitude and variable current amplitude experiments, control 'catch trials', in which no stimulation was delivered, were randomly interleaved with stimulation trials. Electrical stimuli consisted of 25–550 ms trains of biphasic current pulses at 400 Hz (0.4 ms per phase; 2.5 ms between phases)¹⁴. Current amplitudes varied between 10 and 100 μ A. Microstimulation was applied only in the 'undirected' condition (see below, $n = 5$ birds) and never evoked vocalization in quiescent animals.

Lesions

Electrolytic lesions of LMAN were performed and evaluated as previously described⁴. The percentage of LMAN that was removed bilaterally ranged from 50% to 100%.

Behavioural analysis

Undirected song was recorded when the bird was alone. To elicit directed song, one or more female zebra finches were presented in a separate cage. Each female presentation lasted ≤ 2 min, and songs were classified as directed only when the male faced the female(s). Female presentations were interleaved with bouts of undirected singing.

Analysis of neural signals

Analysis of singing-related activity in LMAN was performed as previously described^{12,13}. Briefly, rectified, smoothed neural activity waveforms were aligned using a template for the amplitude envelope of the bird's motif. Both the mean activity level and the coefficient of variation of activity across motif renditions was calculated.

Song analysis

To characterize differences in syllable structure, we measured fundamental frequency. For a particular syllable, we calculated the autocorrelation of a segment of the sound waveform that has constant frequency components (median segment: $\sim 50\%$ of the total syllable duration; range: 20–90%). The fundamental frequency was defined as the distance, in frequency, between the zero-offset peak and the highest peak in the autocorrelation function. To improve the resolution of the estimates, we performed a parabolic interpolation of the autocorrelation function peak.

This algorithm was applied to syllables with clear harmonic structure and a well-defined fundamental frequency or a high frequency, band-limited element. In stimulation experiments, fundamental frequency was measured for at least eight renditions in each condition. To investigate natural differences between directed and undirected songs, fundamental frequency was measured for at least 19 renditions in each context.

To characterize the effects of microstimulation on amplitude, the sound waveform was filtered between 300 and 8,000 Hz, rectified, and smoothed with a 2-ms window. Syllables were segmented using an amplitude threshold, and amplitude was quantified by measuring the area under the rectified waveform from syllable onset to syllable offset.

Statistics

Comparisons of effects across different experimental conditions were made using the

non-parametric Mann–Whitney *U*-test (for within-syllable changes) and paired sign test (for group changes). Comparisons of variability of fundamental frequency in different experimental conditions were made using the *F*-test for equality of variance. In all cases, the minimum significance level was set at $P < 0.05$.

Received 3 August; accepted 21 October 2004; doi:10.1038/nature03127.

- Graybiel, A. M., Aosaki, T., Flaherty, A. W. & Kimura, M. The basal ganglia and adaptive motor control. *Science* **265**, 1826–1831 (1994).
- Bottjer, S. W., Miesner, E. A. & Arnold, A. P. Forebrain lesions disrupt development but not maintenance of song in passerine birds. *Science* **224**, 901–903 (1984).
- Scharff, C. & Nottebohm, F. A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: implications for vocal learning. *J. Neurosci.* **11**, 2896–2913 (1991).
- Brainard, M. S. & Doupe, A. J. Interruption of a basal ganglia–forebrain circuit prevents plasticity of learned vocalizations. *Nature* **404**, 762–766 (2000).
- Williams, H. & Mehta, N. Changes in adult zebra finch song require a forebrain nucleus that is not necessary for song production. *J. Neurobiol.* **39**, 14–28 (1999).
- Tchernichovski, O., Mitra, P. P., Lints, T. & Nottebohm, F. Dynamics of the vocal imitation process: how a zebra finch learns its song. *Science* **291**, 2564–2569 (2001).
- Williams, H., Cynx, J. & Nottebohm, F. Timbre control in zebra finch (*Taeniopygia guttata*) song syllables. *J. Comp. Psychol.* **103**, 366–380 (1989).
- Nottebohm, F., Stokes, T. M. & Leonard, C. M. Central control of song in the canary. *J. Comp. Neurol.* **165**, 457–486 (1976).
- Perkel, D. J. in *Behavioral Neurobiology of Birdsong* (eds Zeigler, H. P. & Marler, P.) 736–748 (New York Academy of Sciences, New York, 2004).
- Troyer, T. W. & Doupe, A. J. An associational model of birdsong sensorimotor learning. II. Temporal hierarchies and the learning of song sequence. *J. Neurophys.* **84**, 1224–1239 (2000).
- Doya, K. & Sejnowski, T. J. in *The New Cognitive Neurosciences* (ed. Gazzaniga, M. S.) 469–482 (MIT Press, Cambridge, Massachusetts, 2000).
- Hessler, N. A. & Doupe, A. J. Social context modulates singing-related neural activity in the songbird forebrain. *Nature Neurosci.* **2**, 209–211 (1999).
- Hessler, N. A. & Doupe, A. J. Singing-related neural activity in a dorsal forebrain–basal ganglia circuit of adult zebra finches. *J. Neurosci.* **19**, 10461–10481 (1999).
- Vu, E. T., Mazurek, M. E. & Kuo, Y. C. Identification of a forebrain motor programming network for the learned song of zebra finches. *J. Neurosci.* **14**, 6924–6934 (1994).
- Vicario, D. S. & Simpson, H. B. Electrical stimulation in forebrain nuclei elicits learned vocal patterns in songbirds. *J. Neurophys.* **73**, 2602–2607 (1995).
- Brumm, H. & Todt, D. Male–male vocal interactions and the adjustment of song amplitude in a territorial bird. *Anim. Behav.* **67**, 281–286 (2004).
- Yu, A. C. & Margoliash, D. Temporal hierarchical control of singing in birds. *Science* **273**, 1871–1875 (1996).
- Johnson, F., Sablan, M. M. & Bottjer, S. W. Topographic organization of a forebrain pathway involved with vocal learning in zebra finches. *J. Comp. Neurol.* **358**, 260–278 (1995).
- Mooney, R. Different subthreshold mechanisms underlie song selectivity in identified HVC neurons of the zebra finch. *J. Neurosci.* **20**, 5420–5436 (2000).
- Jarvis, E. D., Scharff, C., Grossman, M. R., Ramos, J. A. & Nottebohm, F. For whom the bird sings: context-dependent gene expression. *Neuron* **21**, 775–788 (1998).
- Hikosaka, O., Nakamura, K., Sakai, K. & Nakahara, H. Central mechanisms of motor skill learning. *Curr. Opin. Neurobiol.* **12**, 217–222 (2002).
- Miller, E. K. The prefrontal cortex and cognitive control. *Nature Rev. Neurosci.* **1**, 59–65 (2000).
- Troyer, T. W. & Bottjer, S. W. Birdsong: models and mechanisms. *Curr. Opin. Neurobiol.* **11**, 721–726 (2001).
- Mooney, R. & Konishi, M. Two distinct inputs to an avian song nucleus activate different glutamate receptor subtypes on individual neurons. *Proc. Natl Acad. Sci. USA* **88**, 4075–4079 (1991).
- Kittelberger, J. M. & Mooney, R. Lesions of an avian forebrain nucleus that disrupt song development alter synaptic connectivity and transmission in the vocal premotor pathway. *J. Neurosci.* **19**, 9385–9398 (1999).
- Stark, L. L. & Perkel, D. J. Two-stage input-specific synaptic maturation in a nucleus essential for vocal production in the zebra finch. *J. Neurosci.* **19**, 9107–9116 (1999).
- Komatsu, H. & Wurtz, R. H. Modulation of pursuit eye movements by stimulation of cortical areas MT and MST. *J. Neurophys.* **62**, 31–47 (1989).
- Tanaka, M. & Lisberger, S. G. Regulation of the gain of visually guided smooth-pursuit eye movements by frontal cortex. *Nature* **409**, 191–194 (2001).
- Canales, J. J. & Graybiel, A. M. A measure of striatal function predicts motor stereotypy. *Nature Neurosci.* **3**, 377–383 (2000).
- Matsumoto, N., Hanakawa, T., Maki, S., Graybiel, A. M. & Kimura, M. Nigrostriatal dopamine system in learning to perform sequential motor tasks in a predictive manner. *J. Neurophys.* **82**, 978–997 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank N. Hessler for a portion of the data in Fig. 3e. We thank A. Basbaum, S. Lisberger, J. Sakata and B. Wright for helpful comments on this manuscript and A. Arteseros and K. McManaway for technical assistance. This work was supported by an HHMI Predoctoral Fellowship (M.H.K.), the MacArthur Foundation, the Steven and Michele Kirsch Foundation, NARSAD and NIH (A.J.D.), and the HHMI Biomedical Research Support Program grant, the McKnight Foundation, the Klingenstein Fund, a Searle Scholars Award and NIH (M.S.B.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.H.K. (mimi@phy.ucsf.edu).

..... **Ultrabithorax is required for membranous wing identity in the beetle *Tribolium castaneum***

Yoshinori Tomoyasu¹, Scott R. Wheeler^{2*} & Robin E. Denell¹

¹Division of Biology, Chalmers Hall, Kansas State University, Manhattan, Kansas 66506, USA

²Department of Genetics, Washington University School of Medicine, 4566 Scott Avenue, St Louis, Missouri 63110, USA

* Present address: Program in Molecular Biology and Biotechnology, The University of North Carolina at Chapel Hill, Campus Box 3280 - Fordham Hall, Chapel Hill, North Carolina 27599-3280, USA

The two pairs of wings that are characteristic of ancestral pterygotes (winged insects) have often undergone evolutionary modification. In the fruitfly, *Drosophila melanogaster*, differences between the membranous forewings and the modified hindwings (halteres) depend on the Hox gene *Ultrabithorax* (*Ubx*). The *Drosophila* forewings develop without Hox input, while *Ubx* represses genes that are important for wing development, promoting haltere identity^{1,2}. However, the idea that Hox input is important to the morphologically specialized wing derivatives such as halteres, and not the more ancestral wings, requires examination in other insect orders. In beetles, such as *Tribolium castaneum*, it is the forewings that are modified (to form elytra), while the hindwings retain a morphologically more ancestral identity. Here we show that in this beetle *Ubx* 'despecializes' the hindwings, which are transformed to elytra when the gene is knocked down. We also show evidence that elytra result from a Hox-free state, despite their diverged morphology. *Ubx* function in the hindwing seems necessary for a change in the expression of *spalt*, *iroquois* and *achaete-scute* homologues from elytron-like to more typical wing-like patterns. This counter-acting effect of *Ubx* in beetle hindwings represents a previously unknown mode of wing diversification in insects.

Many modern insects have wings on their second (T2) and third (T3) thoracic segments. Wing morphology often differs greatly between species, and sometimes between forewing and hindwing in the same species. In *Drosophila*, the forewing is used for flight, while the hindwing (haltere) is highly reduced and used only for balance (Fig. 1a). *Ubx* promotes haltere identity by repressing expression of some wing genes, including those of the *spalt* (*sal*) complex² (Fig. 1a), but not others such as *optomotor blind* (*omb*; *bifid*, *bi* in Flybase)². Removing *Ubx* function causes the transformation of haltere to forewing¹ (typically referred to simply as 'wing'). In contrast, the forewing is thought to be a Hox-free state, because inactivating or overexpressing *Antennapedia* (*Antp*), the only Hox gene expressed in the forewing, has almost no effect on wing morphology^{3,4}. No wings develop on T1 or the abdominal segments, because *Sex combs reduced* (*Scr*), *Ubx* and *abdominal-A* (*abd-A*) repress wing development in these segments³ (Fig. 1a). Despite the divergence of hindwing morphology between dipterans (flies) and lepidopterans (butterflies and moths), *Ubx* also regulates hindwing identity in the butterfly *Precis coenia*, albeit by regulating a different set of target genes than those in *Drosophila*^{5,6}. Weatherbee *et al.*⁵ proposed that diversification of wing morphology among insects was achieved both by modification of a basic wing-gene network that controls both fore- and hindwing development in a species-specific and *Ubx*-independent manner, and by the divergence of *Ubx*-regulated target genes (rather than by changes in *Ubx* expression) in the hindwing.

Applying these models to beetle wing development is confusing, as the situation in beetles is the opposite to that in *Drosophila*: the T2 segment bears sclerotized elytra (wing covers) (Fig. 1b, c), whereas

the T3 segment develops more typical membranous flight wings (hereafter referred to as 'hindwings') (Fig. 1e). One possible explanation is that *Ubx* is expressed differently in beetles. However, detection of Ultrathorax (Utx), the protein product of the *Tribolium* *Ubx* orthologue, by monoclonal antibody FP6.86 (ref. 7) shows that Utx, like its *Drosophila* counterpart, is expressed in T3 but not in T2 imaginal discs (Fig. 1f, g). This indicates that the function of *Ubx* and/or other Hox genes could be different in beetles.

To examine the function of Hox genes in beetle wing differentiation, we analysed the adult phenotypes of Hox gene mutants in *Tribolium*. Loss-of-function mutants of *Cephalothorax* (*Cx*), the *Tribolium* counterpart of *Scr* (ref. 8), have additional elytron-like tissue on the pronotum (T1)^{9,10} (Fig. 1h), suggesting that *Scr/Cx* function in repressing T1 wing development is conserved between fly and beetle. *Scr/Cx* does not seem to have any further effect on wing development, as elytron and hindwing morphology is normal in these mutants (data not shown). Mutation of the *Antp* orthologue, *prothoraxless* (*ptl*)^{10–12}, causes reduction of the pronotum¹⁰

(Fig. 1i) but has no visible effect on hindwings (data not shown) or elytra (Fig. 1i). However, it is unclear whether the pronotum phenotype results from loss- or gain-of-function. The situation is similar for *Ubx/Utx* mutants, as some mutants (for example, *Utx*¹) that have a loss-of-function phenotype in embryogenesis, seem to have a gain-of-function phenotype affecting elytra at the adult stage¹³.

To determine the Hox gene functions during wing differentiation more clearly in this beetle, we have established a larval RNA interference (RNAi) technique to knock down gene function during imaginal development¹⁴ (see Methods). An enhancer-trap line (*pu11*), which expresses green fluorescent protein (GFP) in hindwing and elytron discs¹⁵ (Fig. 2a, f), was used to visualize hindwing and elytron development in larvae. Injection of *Scr/Cx* double-stranded (ds)RNA at a late larval stage (sixth to eighth instar) induces additional GFP-positive tissues in the T1 segment of larvae (Fig. 2b) and ectopic elytra on the pupal and adult pronotum (Fig. 2d; Supplementary Fig. 1). This phenotype is consistent with that of existing *Cx* mutants (Fig. 1h), indicating that RNAi is working properly in beetle larvae and pupae. We then used larval RNAi to determine the function of *Ubx/Utx* during hindwing/elytron development. *Ubx/Utx* RNAi induced complete transformation of hindwing to elytron (Fig. 2c, e). The transformed elytra

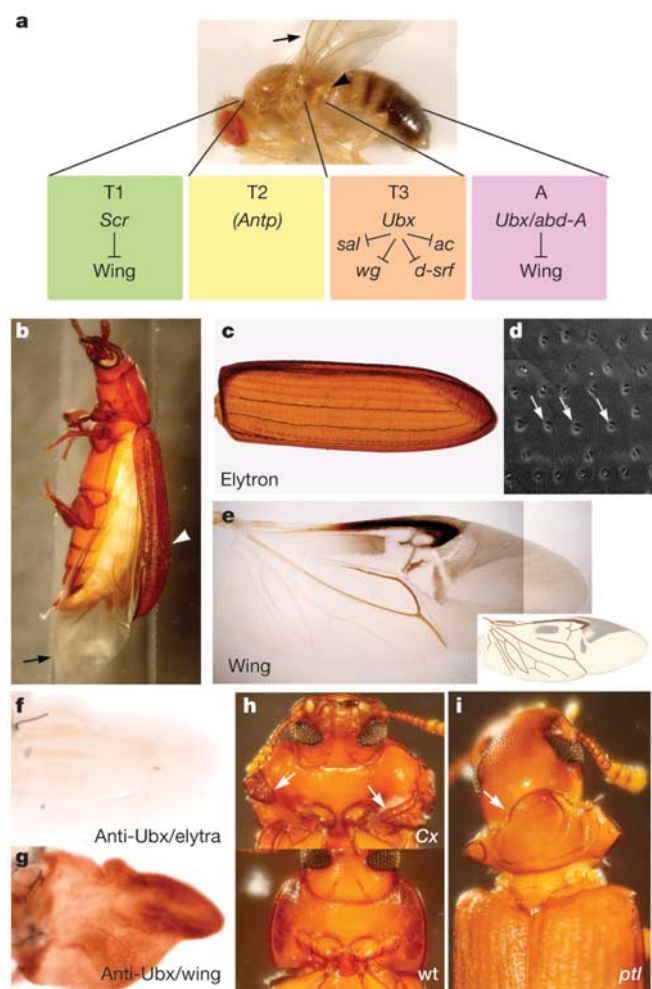


Figure 1 Function of Hox genes in fore- and hindwing differentiation in insects. **a**, A model for fore- and hindwing differentiation in *Drosophila*. Wing (arrow) and haltere (arrowhead) are indicated. *wg*, wingless; *d-srf*, *Drosophila* serum response factor. Adult (**b**), elytron (arrowhead in **b**, **c**) and its scanning electron microscopy image (**d**), and hindwing (arrow in **b**, **e**) of *Tribolium castaneum*. Elytra are covered by numerous sensory bristles (**d**, indicated by arrows). **f**, **g**, *Ubx/Utx* expression in elytra (**f**) and hindwing disc (**g**) monitored by FP6.86. **h**, Ventral view of *Cx*⁶/*Cx*^{apt}. Ectopic tissue on the pronotum is indicated by arrows. Bottom: wild type. **i**, *ptl*^{D60}/*ptl*^{D2}. Arrow indicates the reduced pronotum.

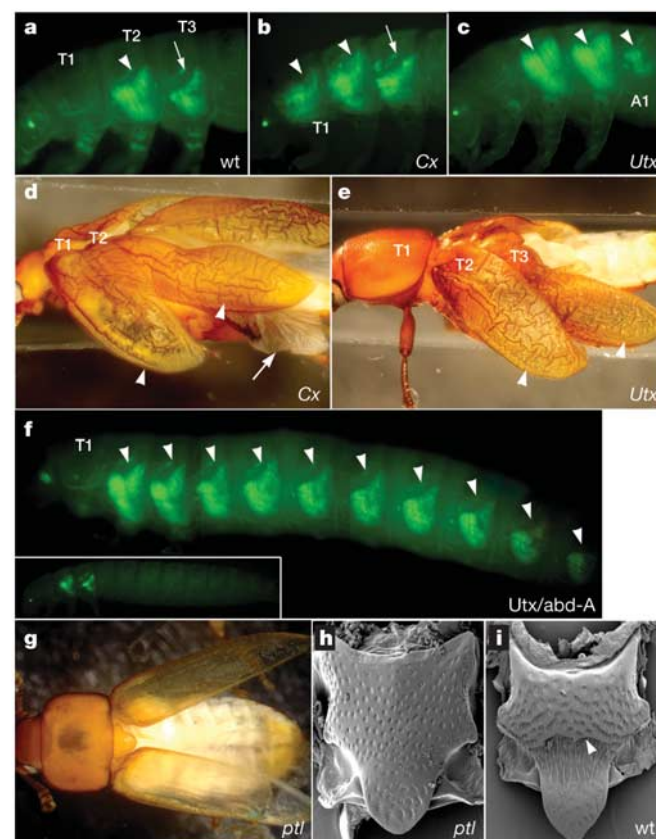


Figure 2 RNAi phenotypes of Hox genes in *Tribolium*. Elytron and wing are denoted by arrowheads and arrows, respectively. **a**, *pu11* larva. GFP is expressed in elytron and hindwing discs. *Scr/Cx* RNAi induces ectopic elytra in the T1 segment (**b**, **d**). *Ubx/Utx* RNAi larva (**c**) and adult (**e**). Hindwings are transformed into elytra (**c**, **e**). A1 has ectopic GFP-expressing tissue (**c**). **f**, *Ubx/Utx* and *abd-A* double RNAi larva, which has 20 elytron discs from T2 to A8. Inset shows full-length of *pu11* larva. *Antp/ptl* RNAi adult (**g**) and mesonotum (**h**). The scutocutellar ridge (arrowhead in **i**) is missing and bristle pattern is uniform in *Antp/ptl* RNAi mesonotum (**h**, compare with **i**, wild type). Anterior is to the left in all images, except for the scanning electron microscopy images (**h**, **i**) in which anterior is up.

often appear almost transparent because *Ubx/Utx* RNAi beetles die before the elytra are sclerotized and pigmented. However, their morphology (for example, sensory structures and vein patterning) is clearly specific to the elytra (Supplementary Fig. 1). *Ubx/Utx* RNAi also affects the first abdominal (A1) segment, inducing small ectopic GFP-expressing regions in larvae (Fig. 2c) and cylindrical projections in adults (Supplementary Fig. 2a), which presumably represent incompletely developed elytra. In addition, dorsal structures of the T3 and A1 segments are transformed to T2 mesonotum (Supplementary Fig. 2a). Double RNAi using *Ubx/Utx* and *Abdominal (A)*, the *abd-A* orthologue, also induces transformation of the hindwing discs to elytra, accompanied by the development of presumptive elytra in all larval abdominal segments (Fig. 2f); these larvae did not survive to the pupal stage. These phenotypes parallel the results in *Drosophila* with respect to the segments affected, but are unexpected on the basis of the regulatory paradigm indicated by *Drosophila* studies. Unlike *Drosophila Ubx*, which normally modifies membranous wing development to produce halteres, *Ubx/Utx* in this beetle seems to promote membranous hindwing development by repressing elytron identity.

In contrast to the marked phenotypes of *Ubx/Utx* and *abd-A/A* RNAi, *Antp/ptl* RNAi induces a rather mild adult phenotype. Several furrow structures (or sutures) on the mesonotum (the median ridge and scutoscuteellar ridge¹⁶) are missing and the microchaete pattern on this structure is more uniform than in the wild type (Fig. 2g, h; compare with Fig. 2i; see Supplementary Fig. 2b, c). However, there is no identifiable effect on the hindwings or elytra (Fig. 2g). These observations indicate that *Antp/ptl* has no obvious function in wing differentiation. These data therefore indicate that the elytra represent a Hox-free state, despite their derived morphology, and that *Ubx/Utx* promotes membranous wing identity in the hindwing of beetles.

To understand how *Ubx/Utx* regulates hindwing development,

we analysed the expression patterns of more than 15 wing genes in *Tribolium* hindwing and elytron imaginal discs. Most genes are expressed similarly in elytron and hindwing (Y.T., manuscript in preparation). However, *Tribolium* homologues of *sal*, *iroquois (iro)* complex genes and *achaete-scute (ac-sc)* are expressed differently in T2 versus T3 imaginal discs (Fig. 3). In the *Drosophila* wing disc, *sal* is induced by Decapentaplegic (Dpp) signalling along the anterior/posterior compartment boundary¹⁷. A *Tribolium sal* homologue (*Tc-sal*) is expressed in the distal part of the hindwing disc during the last larval stage (Fig. 3f). However, the only *Tc-sal* expression in the elytron disc is faint posterior marginal expression (Fig. 3b inset). In contrast, *Tc-dpp* (ref. 18) and *Tc-omb* (a homologue of another gene regulated by Dpp in *Drosophila*)¹⁹, are expressed similarly in both the hindwing and elytron discs (Fig. 3a, e; data not shown). *Tc-iro* is expressed in two short stripes in the hindwing disc of *Tribolium* (Fig. 3g). Double staining with an antibody against Engrailed (Supplementary Fig. 3) suggests that these stripes correspond to *iro* expression in the L3 and L5 veins of the *Drosophila* wing disc²⁰. In the elytron disc, *Tc-iro* is expressed in sensory organ precursor cells, but no stripe expression is observed (Fig. 3c). We also noticed that the *Tc-ac-sc-homologue (Tc-ASH)*²¹ is expressed in stripes across the entire elytron disc (Fig. 3d), but only a few *Tc-ASH*-positive regions are seen in hindwing discs (Fig. 3h). *ac-sc* is known to be important for sensory organ development²², and the observed expression is consistent with the sensory organ pattern of these structures. The *Tc-sal* expression at the posterior margin of the elytron disc (Fig. 3c) is interesting as this region corresponds to the anal marginal area for articulation¹⁶, a less sclerotized band of tissue in adult elytra. *Tc-sal* expression might be important for preventing sclerotization in beetle elytron and hindwing. It is also intriguing that both *iro* and *sal* are important for vein positioning in the *Drosophila* wing disc^{17,20,23,24}. Modification of the gene network for wing vein formation might be important for hindwing/elytron

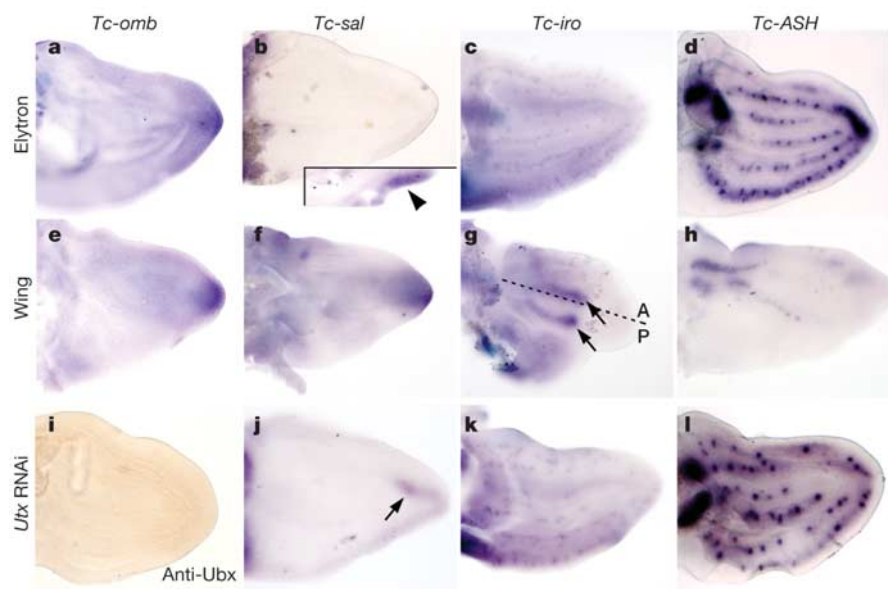


Figure 3 Wing genes' expression in hindwing and elytron disc. *Tc-omb* is expressed similarly in both elytron and hindwing discs (a, e), however, *Tc-sal* expression is absent in the elytron disc (b, f), except for faint expression along the posterior margin, which can be seen after longer staining reaction (arrowhead in inset in b). Two stripes of *Tc-iro* expression are observed in hindwing (arrows in g), but not elytron discs (c). Dashed line indicates the anterior/posterior border. *Tc-iro* is expressed in sensory organ precursor cells in elytra (c). *Tc-ASH* expression is detected in several stripes in elytron discs (d), but is largely absent in hindwing discs (h). *Ubx/Utx* (i), *Tc-sal* (j), *Tc-iro* (k) and *Tc-ASH* (l)

expression in *Ubx/Utx* RNAi hindwings. *Ubx/Utx* is not detected by FP6.86, indicating that RNAi is working properly (i). Distal expression of *Tc-sal* and two *Tc-iro* stripes are missing in these discs (j, k). Residual *Tc-sal* expression is sometimes observed (arrow in j). *Tc-ASH* is ectopically expressed in *Ubx/Utx* RNAi discs, although the stripes are sometimes disoriented (l). These expression patterns in *Ubx/Utx* RNAi hindwing discs are more similar to those in elytron discs. Discs are shown with anterior up and distal to the right.

differentiation. To determine whether *Ubx/Utx* regulates the expression of these genes, we analysed the expression of *Tc-sal*, *Tc-iro* and *Tc-ASH* in *Ubx/Utx* RNAi hindwing discs, and found that expression of all three genes is altered. *Tc-sal* distal expression is greatly reduced (sometimes completely missing) and faint posterior margin expression is observed instead (Fig. 3j). The two stripes of *Tc-iro* expression are missing (Fig. 3k), and *Tc-ASH* is ectopically expressed (Fig. 3l). These results indicate that *Ubx/Utx* promotes membranous hindwing identity in part by controlling the expression of these genes in *Tribolium*.

It has been proposed that the dorsal appendages of insects first appeared under no (or lesser) Hox influence³. Eventually, some Hox genes gained influence over the wing-gene network, probably via alterations in the *cis*-regulatory elements of these genes³. *Ubx* is expressed in the hindwings of all insects so far examined^{2,6,25}—Diptera, Lepidoptera, Hymenoptera (ants, bees, wasps) and Coleoptera (beetles)—and modifies their structure to varying degrees. Modification of the wing programme by *Ubx* in Lepidoptera and Hymenoptera, as inferred from morphology, seems to be rather minor, resulting in a hindwing structure similar to forewings. In Diptera, the modification by *Ubx* is more extreme and produces the haltere. Our data demonstrate that in beetles, the wing programme has been significantly altered into an 'elytron programme', and that *Ubx/Utx* represses at least some of these changes, resulting in a more ancestral morphology. That is, *Ubx/Utx* has become indispensable for membranous wing differentiation in the beetle lineage, whereas in other insects, membranous wings can develop with or without *Ubx* expression. It could be argued that the elytra are ancestral (on the basis of their location on T2) and that the beetle hindwing is derived. However, our data show that the gene expression patterns of beetle hindwings are more similar to those of fly forewings than are those of the elytra, indicating that the beetle hindwing represents the more ancestral state. This is further supported by the expression pattern of a *sal* homologue in a Hymenopteran insect²⁵. In this case, *sal* expression in both wings resembles that seen in beetle hindwings. Further genetic analysis of beetles as well as other insect orders is necessary to resolve this issue.

There are several possible scenarios to explain the evolution of an 'elytron programme' and the counteracting function of *Ubx/Utx* in beetles. One explanation is that the wing programme has recruited several elytron genes (such as cuticle thickening genes), and that *Ubx/Utx* represses these genes in hindwings. This might explain the textural modification and other morphological changes of the elytra, although it may be necessary to consider several more scenarios to explain the morphology of this highly specialized appendage. For instance, one or more non-Hox genes (referred to as gene(s) X) could have come to control the expression of wing genes such as *Tc-sal*, *Tc-iro* and *Tc-ASH* in a way that promotes elytron identity. *Ubx/Utx* would then suppress the functions of gene(s) X to prevent elytron differentiation in the hindwing. Alternatively, *Ubx/Utx* may have become integrated into the wing-gene network. That is, some genes in the wing network might have acquired *Ubx/Utx* binding sites in their *cis*-regulatory regions, such that this network was no longer able to induce typical membranous wing in the absence of *Ubx/Utx*. Another possibility is that characteristics of the *Utx* protein itself (such as binding affinity) might have been altered such that *Utx* now regulates genes in the wing-gene network. It is possible that a combination of these scenarios accounts for the unusual function of *Utx* in beetles. Finally, although beetle hindwings seem to be less derived than the elytra, it is obvious that the beetle hindwing has also been evolutionarily modified. Thus, *Ubx/Utx* might also modify hindwing in a manner similar to its function in *Drosophila* and other insects². Identification of gene(s) X and analyses of *cis*-elements for the genes that are expressed differently in elytron and hindwing will be necessary for a detailed understanding of *Ubx* function

in beetle hindwing, as well as insight into the general mechanisms regulating fore- and hindwing differentiation in all winged insects. □

Methods

Beetles

All beetle strains were raised at 30 °C. The following mutants and transgenic beetles were used in this study. *Cx⁶* and *Cx^{ap1}* are *Cx* loss-of-function mutants^{9,10}. *Cx^{ap1}* is a weak *Cx* hypomorphic allele, while *Cx⁶* is a null allele that encodes a truncated version of the *Cx* protein (Rogers and Denell, unpublished data). *ptl^{D2}* and *ptl^{D60}* seem to behave as loss-of-function alleles in the embryonic stage²⁶. The *pu11* enhancer-trap line¹⁵, previously referred to as Pig-23 (ref. 15), was also used for visualizing hindwing and elytron development in the larval stage.

Cloning genes

Fragments of beetle genes were obtained by polymerase chain reaction (PCR) with reverse transcription using degenerate primers targeted to the motifs conserved between fly and human. Extra fragments were also obtained by universal PCR²⁷. The motifs targeted in each protein are as follows. *Sal*: MHYRHTGTERPFC (for two nested primers) and VLQQHIR. *Iro*: KNPYPYTKGEKI and ANARRRLKKENKMTW (for two nested primers). *Omb*: GTEMVITKSGR, YKFHNSRWMVAGKAD and TQLKIDNNPFAKG.

Dissection and staining

In *Tribolium*, hindwing and elytron imaginal discs develop during the last larval stage, in a fashion similar to the mealworm beetle²⁸ (Y.T., manuscript in preparation). Crude dissections of larval discs were performed in phosphate-buffered saline (PBS). These large tissue fragments containing imaginal discs were then pre-fixed in 3.7% formaldehyde/PBS for 5 min. After the removal of unnecessary tissues in PBS, imaginal discs were fixed in 3.7% formaldehyde/PBS for 30 min. Antibody staining was performed by standard procedures. Monoclonal antibody FP6.86 (1:5, gift from R. White), which recognizes an evolutionarily conserved epitope common to both *Ubx* and *abd-A*⁷, was used to monitor *Ubx/Utx* expression, and anti-mouse IgG conjugated to horseradish peroxidase (1:200, Jackson) was used as the secondary antibody. The 4D9 anti-Engrailed antibody developed by C. Goodman was obtained from the Developmental Studies Hybridoma Bank (DSHB) (developed under the auspices of the NICHD and maintained by The University of Iowa, Department of Biological Sciences). *In situ* hybridization was performed using standard procedures.

dsRNA synthesis and injection

dsRNA was synthesized using the Ambion MEGAscript high yield transcription kit. The sizes of dsRNA used in this work are as follows. *Cx*: 930 bp; *ptl*: 538 bp; *Utx*: 450 bp; *abd-A*: 621 bp. We injected 0.4–0.5 µg of dsRNA into last (or penultimate) instar *pu11* larvae just before the onset of GFP expression in the hindwing and elytron disc (selected by GFP expression), and demonstrated that RNAi works even in the larval stage of *Tribolium* (larval RNAi)¹⁴. Injected larvae were kept at 30 °C until the proper stage for analysis.

Image processing

Some images were processed using Auto-Montage (Syncroscopy), and brightness and contrast of all images were adjusted with Photoshop (Adobe).

Received 12 October; accepted 13 December 2004; doi:10.1038/nature03272.

- Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**, 565–570 (1978).
- Weatherbee, S. D., Halder, G., Kim, J., Hudson, A. & Carroll, S. Ultrabithorax regulates genes at several levels of the wing-patterning hierarchy to shape the development of the *Drosophila* haltere. *Genes Dev.* **12**, 1474–1482 (1998).
- Carroll, S. B., Weatherbee, S. D. & Langeland, J. A. Homeotic genes and the regulation and evolution of insect wing number. *Nature* **375**, 58–61 (1995).
- Struhl, G. Genes controlling segmental specification in the *Drosophila* thorax. *Proc. Natl Acad. Sci. USA* **79**, 7380–7384 (1982).
- Weatherbee, S. D. et al. Ultrabithorax function in butterfly wings and the evolution of insect wing patterns. *Curr. Biol.* **9**, 109–115 (1999).
- Warren, R. W., Nagy, L., Selegue, J., Gates, J. & Carroll, S. Evolution of homeotic gene regulation and function in flies and butterflies. *Nature* **372**, 458–461 (1994).
- Kelsh, R., Weinzierl, R. O., White, R. A. & Akam, M. Homeotic gene expression in the locust *Schistocerca*: an antibody that detects conserved epitopes in Ultrabithorax and abdominal-A proteins. *Dev. Genet.* **15**, 19–31 (1994).
- Curtis, C. D. et al. Molecular characterization of *Cephalothorax*, the *Tribolium* ortholog of *Sex combs reduced*. *Genesis* **30**, 12–20 (2001).
- Beeman, R. W. A homeotic gene cluster in the red flour beetle. *Nature* **327**, 247–249 (1987).
- Beeman, R. W., Stuart, J. J., Haas, M. S. & Denell, R. E. Genetic analysis of the homeotic gene complex (HOM-C) in the beetle *Tribolium castaneum*. *Dev. Biol.* **133**, 196–209 (1989).
- Brown, S. J. et al. Sequence of the *Tribolium castaneum* homeotic complex: the region corresponding to the *Drosophila melanogaster* Antennapedia complex. *Genetics* **160**, 1067–1074 (2002).
- Stuart, J. J., Brown, S. J., Beeman, R. W. & Denell, R. E. A deficiency of the homeotic complex of the beetle *Tribolium*. *Nature* **350**, 72–74 (1991).
- Bennett, R. L., Brown, S. J. & Denell, R. E. Molecular and genetic analysis of the *Tribolium* Ultrabithorax ortholog, *Ultrathorax*. *Dev. Genes Evol.* **209**, 608–619 (1999).
- Tomoyasu, Y. & Denell, R. E. Larval RNAi in *Tribolium* (Coleoptera) for analyzing adult development. *Dev. Genes Evol.* **214**, 575–578 (2004).
- Lorenzen, M. D. et al. piggyBac-mediated germline transformation in the beetle *Tribolium castaneum*.

Insect Mol. Biol. 12, 433–440 (2003).

16. El-Kifl, A. H. Morphology of the adult *Tribolium confusum* Duv. and its differentiation from *Tribolium (stene) castaneum* Herbst. *Bull. Soc. Foud 1er Entom.* 37, 173–249 (1953).
17. de Celis, J. F., Barrio, R. & Kafatos, F. C. A gene complex acting downstream of *dpp* in *Drosophila* wing morphogenesis. *Nature* 381, 421–424 (1996).
18. Sanchez-Salazar, J. et al. The *Tribolium decapentaplegic* gene is similar in sequence, structure, and expression to the *Drosophila dpp* gene. *Dev. Genes Evol.* 206, 237–246 (1996).
19. Grimm, S. & Pflugfelder, G. O. Control of the gene *optomotor-blind* in *Drosophila* wing development by *decapentaplegic* and *wingless*. *Science* 271, 1601–1604 (1996).
20. Gomez-Skarmeta, J. L., Diez del Corral, R., de la Calle-Mustienes, E., Ferre-Marco, D. & Modolell, J. *aracuan* and *caupolican*, two members of the novel iroquois complex, encode homeoproteins that control proneural and vein-forming genes. *Cell* 85, 95–105 (1996).
21. Wheeler, S. R., Carrico, M. L., Wilson, B. A., Brown, S. J. & Skeath, J. B. The expression and function of the *achaete-scute* genes in *Tribolium castaneum* reveals conservation and variation in neural pattern formation and cell fate specification. *Development* 130, 4373–4381 (2003).
22. Campuzano, S. et al. Molecular genetics of the *achaete-scute* gene complex of *D. melanogaster*. *Cell* 40, 327–338 (1985).
23. Sturtevant, M. A., Bieh, B., Marin, E. & Bier, E. The *spalt* gene links the A/P compartment boundary to a linear adult structure in the *Drosophila* wing. *Development* 124, 21–32 (1997).
24. de Celis, J. F. & Barrio, R. Function of the *spalt/spalt-related* gene complex in positioning the veins in the *Drosophila* wing. *Mech. Dev.* 91, 31–41 (2000).
25. Abouheif, E. & Wray, G. A. Evolution of the gene network underlying wing polyphenism in ants. *Science* 297, 249–252 (2002).
26. Beeman, R. W., Brown, S. J., Stuart, J. J. & Denell, R. E. In *Molecular Insect Science* (eds Hagedorn, H. H., Hildebrandt, J. G., Kidwell, M. G. & Law, J. H.) 21–29 (Plenum, New York, 1990).
27. Beeman, R. W. & Stauth, D. M. Rapid cloning of insect transposon insertion junctions using 'universal' PCR. *Insect Mol. Biol.* 6, 83–88 (1997).
28. Quennedey, A. & Quennedey, B. Morphogenesis of the wing anlagen in the mealworm beetle *Tenebrio molitor* during the last larval instar. *Tissue Cell* 22, 721–740 (1990).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank G. Bucher, M. Weber and M. Klingler for the *pu11* enhancer-trap line, R. White for the FP6.86 antibody, DSHB for the 4D9 antibody and G. Pflugfelder for sharing information about *omb* degenerate primers. We thank K. Leonard for maintaining beetle stocks, T. Shippy for discussion and reading and S. Brown, R. Beeman, S. Haas and all the Manhattan beetle/insect laboratory members for discussion and comments. Y.T. thanks A. Sato and T. Yamaguchi for discussion. This work was supported by the international Human Frontier Science Program Organization (Long-term Fellow) and the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Y.T. (tomoyasu@ksu.edu). The sequences are available in GenBank under the following accession numbers. *Tc-sal*: AY600513; *Tc-iro*: AY600514; *Tc-omb*: AY600516.

Postnatal *Isl1*⁺ cardioblasts enter fully differentiated cardiomyocyte lineages

Karl-Ludwig Laugwitz^{1*}, Alessandra Moretti^{1*}, Jason Lam^{1*}, Peter Gruber³, Yinhong Chen¹, Sarah Woodard¹, Li-Zhu Lin¹, Chen-Leng Cai¹, Min Min Lu¹, Michael Reth⁵, Oleksandr Platoshyn², Jason X.-J. Yuan², Sylvia Evans¹ & Kenneth R. Chien¹

¹Institute of Molecular Medicine and ²Department of Medicine, University of California, San Diego, School of Medicine, La Jolla, California 92093, USA

³Children's Hospital of Philadelphia, Cardiac Center, Philadelphia, Pennsylvania 19104, USA

⁴Cardiovascular Division, Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

⁵Max-Planck Institut für Immunbiologie, Universität Freiburg, Biologie III, Abteilung Molekulare Immunologie, Freiburg 79108, Germany

* These authors contributed equally to this work

The purification, renewal and differentiation of native cardiac progenitors would form a mechanistic underpinning for unravelling steps for cardiac cell lineage formation, and their links to forms of congenital and adult cardiac diseases^{1–3}. Until now there has been little evidence for native cardiac precursor

cells in the postnatal heart⁴. Herein, we report the identification of *Isl1*⁺ cardiac progenitors in postnatal rat, mouse and human myocardium. A cardiac mesenchymal feeder layer allows renewal of the isolated progenitor cells with maintenance of their capability to adopt a fully differentiated cardiomyocyte phenotype. Tamoxifen-inducible Cre/lox technology enables selective marking of this progenitor cell population including its progeny, at a defined time, and purification to relative homogeneity. Co-culture studies with neonatal myocytes indicate that *Isl1*⁺ cells represent authentic, endogenous cardiac progenitors (cardioblasts) that display highly efficient conversion to a mature cardiac phenotype with stable expression of myocytic markers (25%) in the absence of cell fusion, intact Ca²⁺-cycling, and the generation of action potentials. The discovery of native cardioblasts represents a genetically based system to identify steps in cardiac cell lineage formation and maturation in development and disease.

In skeletal muscle, satellite cells are considered specialized endogenous muscle precursors (that is, myoblasts) that are pre-programmed to enter the skeletal muscle lineage^{5,6}. Despite intensive inquiry, there has been no clear evidence for a native cardiac progenitor population with similar characteristics in the myocardium. Various cell surface markers have been used to purify precursor cells from heart muscle including c-kit and sca-1 (refs 7, 8), but the precise role of these cells in *in vivo* cardiogenesis is unclear and the efficiency of conversion to fully differentiated myocytes, in the absence of fusion, is relatively low^{9–11}. Taking advantage of a developmental lineage marker for undifferentiated cardiogenic precursor cells as a requirement for a heart-specific origin, we have identified in the postnatal heart a novel cardiac cell type. The LIM-homeodomain transcription factor *Isl1* (*Isl1*) marks a cell population that makes a substantial contribution to the embryonic heart, comprising most cells in the right ventricle, both atria, the outflow tract and also specific regions of the left ventricle¹².

A subset of *Isl1*⁺ undifferentiated precursors remains embedded in the embryonic mouse heart after its formation and their number decreases progressively from embryonic day 12.5 (ED12.5) to ED18.5 (Fig. 1a–d). After birth, relatively few *Isl1*⁺ cardioblasts were still detectable, averaging 500 to 600 in the myocardium of a 1–5-day-old rat (Fig. 1e, f). Their organ distribution matched the defined contributions of *Isl1*⁺ embryonic precursors (Fig. 1i), suggesting these cells are developmental remnants of the fetal progenitor population. Clusters of *Isl1*⁺ cardioblasts were observed in both atria, whereas in the ventricles they occurred mostly as single cells (Fig. 1e, f). The localization of *Isl1*⁺ progenitors in specific cardiac segments (atrial muscle wall, intra-atrial septum, conus muscle, right ventricle) was conserved among diverse species: mouse, rat and human (Fig. 1g, h; Supplementary Table 1).

Using conditional genetic marking techniques in the mouse, we performed Cre-recombinase-triggered cell lineage tracing experiments to irreversibly mark *Isl1*-expressing cells as well as their differentiated progeny during embryonic development (Fig. 2). *Isl1*-IRES-Cre mice¹³ were crossed into the conditional Cre reporter strain R26R¹⁴, in which Cre-mediated removal of a stop sequence results in the ubiquitous expression of the lacZ gene under the control of the endogenous Rosa26 promoter (Fig. 2a). In neonatal mice bearing both alleles a high proportion of right ventricular myocardium expressed β-galactosidase (β-gal) detected by 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal) staining (Fig. 2a). Around 30–40% of cardiac myocytes isolated from double heterozygous hearts stained positive for X-gal and displayed co-expression of β-gal and sarcomeric α-actinin (Fig. 2b), demonstrating that a significant proportion of myocytes originate from *Isl1*⁺ cardiac progenitors.

To achieve temporal and spatial control of Cre expression

required for tracing the fate of $isl1^{+}$ progenitors resident in the postnatal heart, we generated mice that harbour a knockin of the tamoxifen-dependent Cre recombinase ($isl1$ -mER-Cre-mER) into the $isl1$ locus (Fig. 3). The presence of two mutated oestrogen-receptors (mERs) results in the sequestration of Cre in the cytoplasm¹⁵. Tamoxifen induces a rapid nuclear translocation of the mER-Cre-mER protein, which permits Cre-mediated recombination^{16,17}. As a readout for Cre activity, we again used the reporter strain R26R¹⁴. Thus, application of tamoxifen to double heterozygous $isl1$ -mER-Cre-mER/R26R mice leads to β -gal expression in cells expressing $isl1$ at the time of exposure (Fig. 3a). Experiments with $isl1$ -mER-Cre-mER/R26R mice were performed to assess the specificity of cell labelling. β -gal⁺ cells are found only within the looping heart in double heterozygous animals after tamoxifen

injection at ED7.5 and X-gal staining at ED9.5 (Fig. 3b). Double immunofluorescent analysis confirmed the co-expression of $isl1$ and Cre in undifferentiated cardiac progenitors and demonstrated that nuclear or cytosolic Cre localization was dependent on tamoxifen exposure (Fig. 3c). At ED12, sections of right ventricle and atrial tissue, heart regions that have been shown to originate from $isl1^{+}$ precursors¹², exhibited expression of α -myosin heavy chain but were negative for $isl1$ and Cre (Fig. 3d). Thus, tamoxifen injection leads to a highly specific marking of $isl1$ -expressing progenitors and Cre protein is not detectable ~ 54 h after the promoter has been inactivated. After birth, β -gal⁺ cells can be detected with highest frequency in the conus muscle, the right atrium and the right ventricle by X-gal stain in $isl1$ -mER-Cre-mER/R26R double heterozygous neonatal hearts after tamoxifen injection at ED17

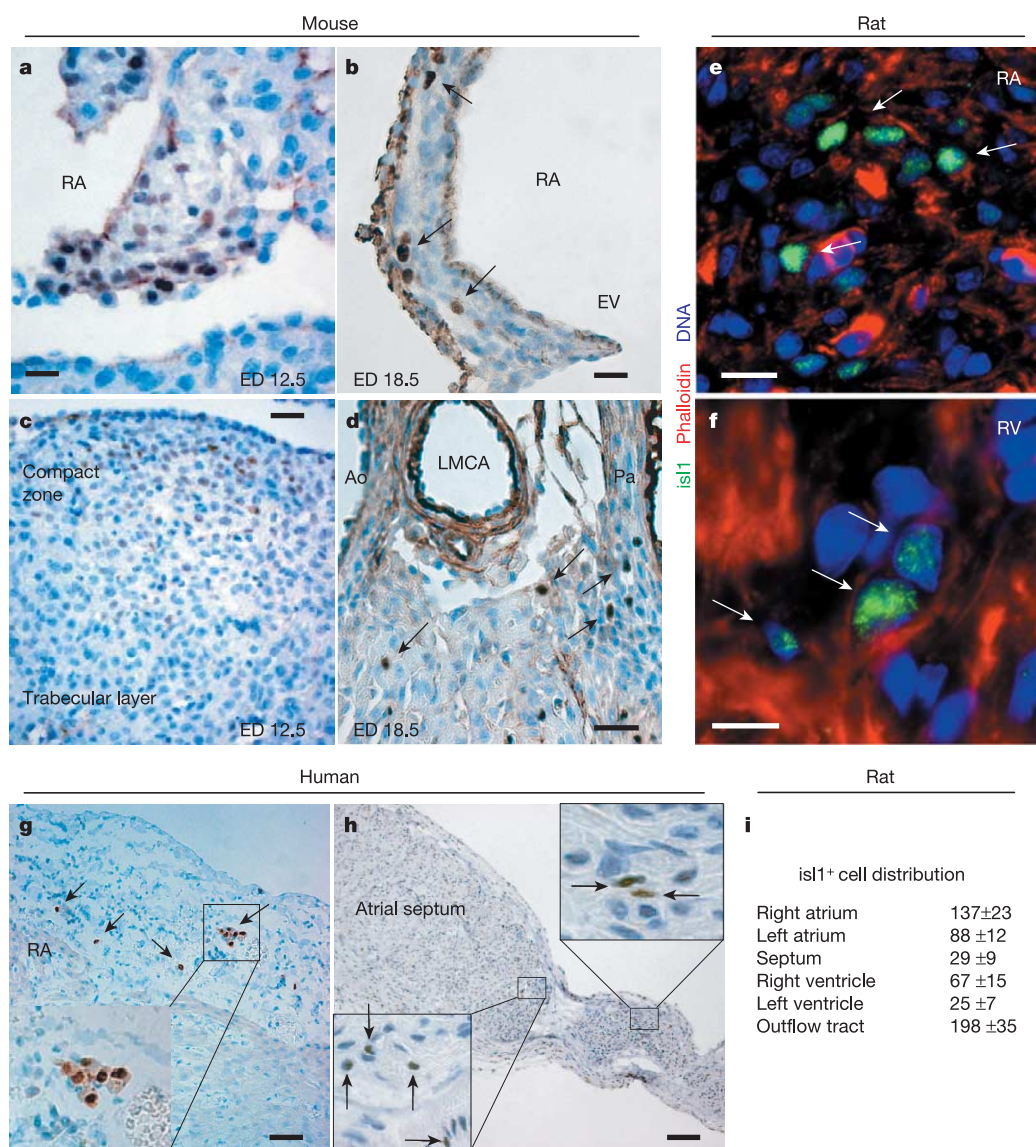


Figure 1 $Isl1^{+}$ progenitors in the late embryonic and postnatal heart. **a-d**, $Isl1^{+}$ progenitors in mouse sections at ED12.5 (**a**, atrial septum; **c**, ventricular tissue) and 18.5 (**b**, free atrial wall; **d**, ventricular arterial region). Scale bar, 20 μ m (**a, b, d**), 50 μ m (**c**). EV, eustachian valve; RA, right atrium; Ao, aorta; LMCA, left main coronary artery; Pa, pulmonary artery. **e, f**, Cluster of $isl1^{+}$ cardiac precursors embedded in right atrial tissue and in the right ventricle of a postnatal day 1 rat heart. Scale bars, 10 μ m (**e**), 15 μ m (**f**).

g, h, $Isl1^{+}$ progenitors in human right atrial tissue from an 8-day-old patient and intra-atrial septum from a 2-day-old patient. Scale bars, 75 μ m (**g**), 150 μ m (**h**).

i, Quantification and localization of $isl1^{+}$ cells in postnatal day 1 rat hearts. Mean values $\pm$ s.e.m. from three hearts. Arrows designate $isl1^{+}$ cells in different species and insets represent a magnification of the areas of interest.

(Fig. 3e), and represent a rare subset of cells in the adult right atrium and outflow tract (Supplementary Fig. 1).

Following cell isolation from postnatal *isl1*-mER-Cre-mER/R26R mouse hearts and reporter gene induction *in vitro*, most β -gal⁺ cells were found in the cardiac mesenchymal cell (CMC) fraction (Fig. 4a). Mesenchymal cells isolated from postnatal rat hearts also contain *isl1*⁺ progenitors (Supplementary Fig. 2). The CMC culture environment maintains *isl1* expression in the progenitor population and promotes their expansion in culture without differentiation (Fig. 4b and Supplementary Fig. 2).

β -gal⁺ cells can be purified from the mesenchymal cells using fluorescence-activated cell sorting (FACS) after labelling with a fluorogenic, lipophilic β -gal substrate (C₁₂FDG) (Fig. 4c, d). *Isl1*⁺ progenitors were isolated as a distinct population presenting seven to ten times higher C₁₂FDG fluorescence than background at a purity of 90–95%. They corresponded to 0.5% of the mesenchymal cell fraction when FACS-sorted during an early stage in culture (4–6 days) and reached ~8% after 15–18 days culture (Fig. 4c, d). As phenotypic characteristics of an undifferentiated state, β -gal⁺ precursors expressed *isl1* and early specification markers for cardiac mesoderm, *Nkx2.5* and *GATA-4*, while lacking transcripts of mature myocytes (Fig. 4e).

FACS analysis of the FDG⁺ cell fraction, expanded *in vitro* over 10 days, revealed negative staining for *sca-1* and no correlation between β -gal expression and the efflux of the DNA-binding dye Hoechst 33342, demonstrating that *isl1*⁺ cardioblasts are distinct from cardiac side population cells (Fig. 4f–h and Supplementary Table 2)^{8,18}. In addition, *isl1*⁺ cells did not express *c-kit*.

After exposure to 4-hydroxytamoxifen (4-OH-TM) *in vitro*, a rare subset of β -gal⁺ cells coexpressing cardiac troponin T was observed at day 4 ($n = 51 \pm 12$), day 7 ($n = 81 \pm 29$) and day 10 ($n = 157 \pm 10$) in the myocyte fraction (Fig. 4i, j), resulting in ~75% of the total β -gal⁺ cells. The fact that *isl1* is repressed as

soon as the cells acquire a differentiated phenotype¹² and the lack of β -gal expression in the absence of 4-OH-TM (Fig. 3b) strongly suggest that *isl1*⁺ cardioblasts can spontaneously acquire myocytic characteristics when dispersed and cultured in the presence of differentiated myocytes.

To assess the potential of β -gal⁺ progenitors to adopt a myocytic phenotype, we performed co-culture experiments of FACS-sorted precursors with neonatal cardiac myocytes. After 3 days in co-culture, β -gal⁺ cells were found within the myocytic syncytium and expressed cardiac specific proteins, that is, α -sarcomeric actinin or troponin T in a partially or completely organized pattern (Fig. 4k, l). Moreover, a subset of β -gal⁺ cardioblasts showed contractile activity and intact electromechanical coupling to neighbouring cells, as documented by expression of the gap-junctional protein connexin 43 (data not shown). The differentiation capacity of β -gal⁺ precursors did not significantly differ between cells maintained in culture for only a few days (early stage) and cells expanded for 15–18 days (late stage) (Fig. 4l).

We addressed the role of cell fusion in the progenitor-to-cardiac myocyte transition, by performing co-culture experiments in which neonatal myocytes were pre-fixed with paraformaldehyde to stabilize cellular components and render the cells refractory to cell fusion¹⁹. Paraformaldehyde treatment disrupted the exclusion of propidium iodide by the myocytes, indicating that these cells were non-viable (Fig. 4m). FACS-sorted β -gal⁺ and β -gal[−] cells were labelled with semiconductor red-fluorescent nanocrystals (QD₆₅₅)²⁰ and co-cultured with pre-fixed myocytes in the presence of conditioned medium from viable myocytes. Under these conditions, within 5 days around 30% of the QD₆₅₅-marked β -gal⁺ progenitors started to express α -actinin (Fig. 4m), indicating that cardioblasts enter the myocytic program independently from cell fusion. In contrast, all β -gal[−] cells stained negative for

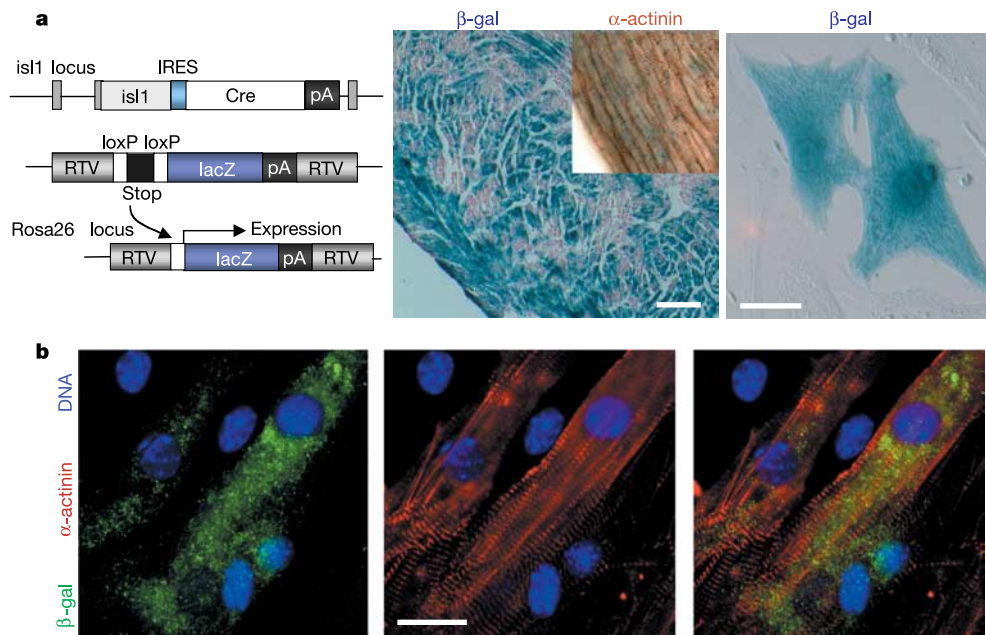


Figure 2 Genetic marking of *isl1*⁺ progenitors and myocytic cell fate. **a**, Mice carry one *isl1*-IRES-Cre allele and one R26R reporter gene. Cre expression catalyses excision of the stop cassette, resulting in selective lacZ expression and genetic marking of *isl1*-expressing cells and their differentiated progeny. RTV, retroviral integration sequence. Shown are β -gal⁺ cardiomyocytes in the right ventricle (middle panel; inset demonstrates

α -actinin expression; scale bar, 180 μ m) and after cell isolation (right panel; scale bar, 20 μ m) from double heterozygous hearts of 4-month- or 1-day-old animals, respectively. **b**, Immunocytochemistry for β -gal and sarcomeric α -actinin in isolated cardiac myocytes from animals carrying both alleles. Scale bar, 15 μ m.

α -actinin, as well as β -gal⁺ precursors when cultured in the absence of pre-fixed myocytes, demonstrating that differentiation is cell-specific and requires secreted and membrane-bound factors (Fig. 4m).

To characterize whether the β -gal⁺ cardioblasts could adopt a

fully differentiated cardiac phenotype, C₁₂FDG-labelled cells in co-culture were analysed by real-time intracellular calcium [Ca²⁺]_i imaging. Differentiated C₁₂FDG⁺ progenitors showed periodic [Ca²⁺]_i oscillations similar to and synchronized with those in adjacent myocytes (Fig. 5a, b; Supplementary Video). Isoproterenol

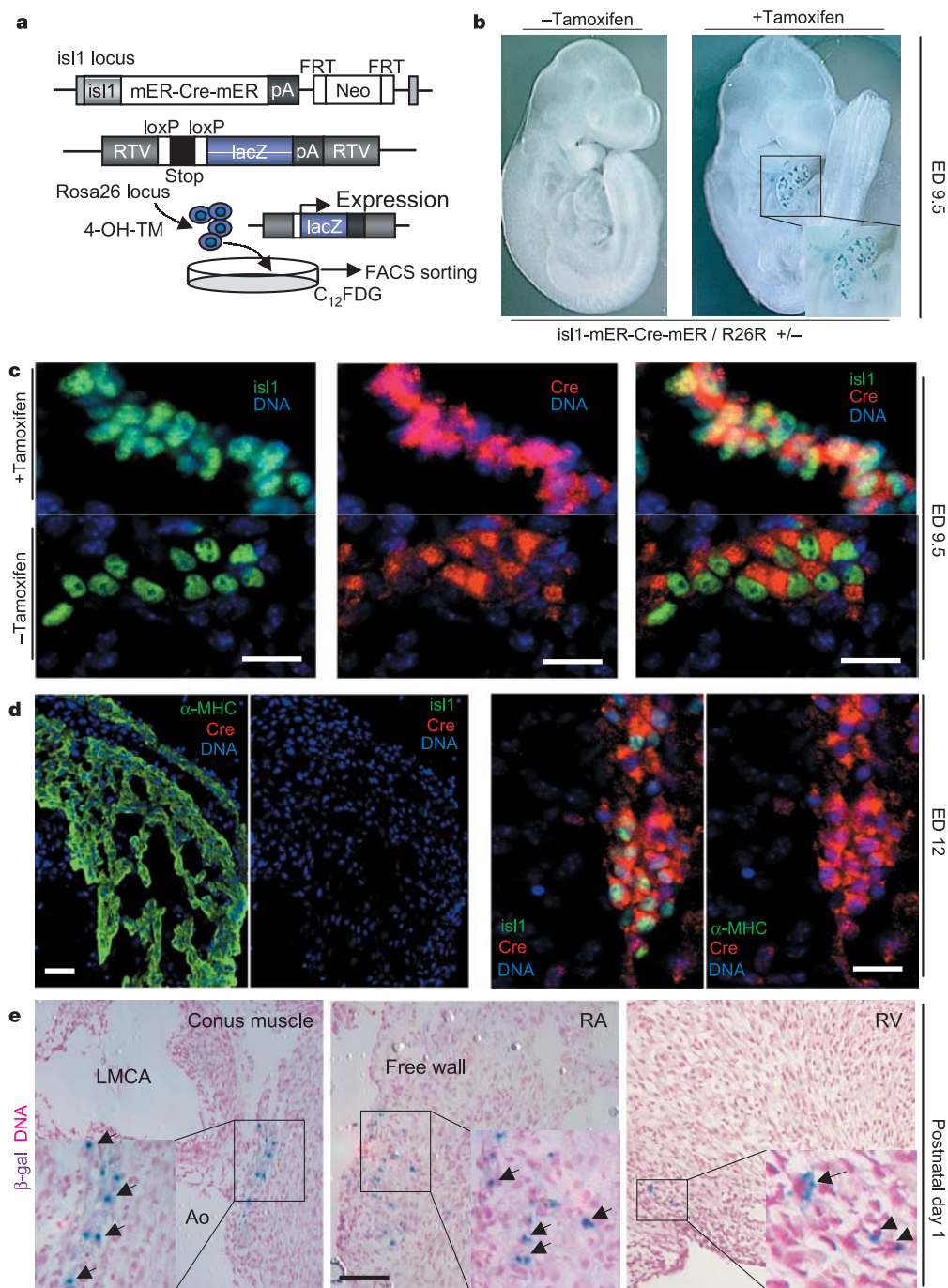


Figure 3 Specificity of recombination and distribution of β -gal⁺ cells in *isl1*-mER-Cre-mER/R26R hearts. **a**, Tamoxifen injection of *isl1*-mER-Cre-mER/R26R double heterozygous mice or administration of 4-OH-TM in culture results in heritable expression of *lacZ*. Labelling with C₁₂FDG allows FACS purification of *isl1*⁺ cells. **b**, X-gal stain in whole-mount double heterozygous embryos with or without tamoxifen injection into pregnant mothers at ED7.5. **c**, Cre localization in *isl1*⁺ cells from *isl1*-mER-Cre-mER/R26R embryos is cytosolic without tamoxifen and nuclear with tamoxifen injection. Scale

bar, 25 μm. **d**, Downregulation of *isl1* and Cre in sequential sections of right ventricular and atrial tissue (left panels; scale bar, 80 μm) but not of dorsal mesocardium (right panels; scale bar, 30 μm). **e**, Sections from hearts of 1-day-old double heterozygous mice injected with tamoxifen at ED17, after X-gal stain and haematoxylin counterstain. LMCA, left main coronary artery; Ao, aorta; RA, right atrium; RV, right ventricle. Scale bar, 150 μm. Arrows designate β -gal⁺ cells and insets represent a magnification of the areas of interest.

stimulation induced a 30% increase of the Ca^{2+} transient amplitude in differentiated $\beta\text{-gal}^+$ cells, suggesting the existence of a functional β -adrenergic signalling pathway²¹ (Fig. 5c). The $[\text{Ca}^{2+}]_i$ oscillations responded to increasing frequencies of external field stimulation (Fig. 5d). Of the C_{12}FDG -labelled cardioblasts in the co-culture 2.3% showed active intracellular Ca^{2+} transients. Functional maturity was confirmed by analysing action potential characteristics in QD_{655} -marked progenitors and neighbouring myocytes. Beating differentiated cardioblasts displayed continuous electrical activity or action potentials characterized by upstroke velocity (dV/dt_{max}), amplitude (APA), duration (APD) and maximum diastolic

potential (MDP) comparable to those of neonatal cardiac myocytes (Fig. 5e, f). Repeated impalements of individual QD_{655} -marked differentiated progenitors revealed reproducible action potentials with distinct morphologies, nodal-like and atrial-like, supporting the concept that these cells are able to differentiate into a distinct subset of cardiac lineages, including conduction system myocytes²².

We demonstrate that the mammalian heart harbours an organ-specific progenitor cell that can be localized, purified, expanded and differentiated into mature cardiac myocytes (Fig. 5g). It will be interesting to investigate whether $\text{isl}1^+$

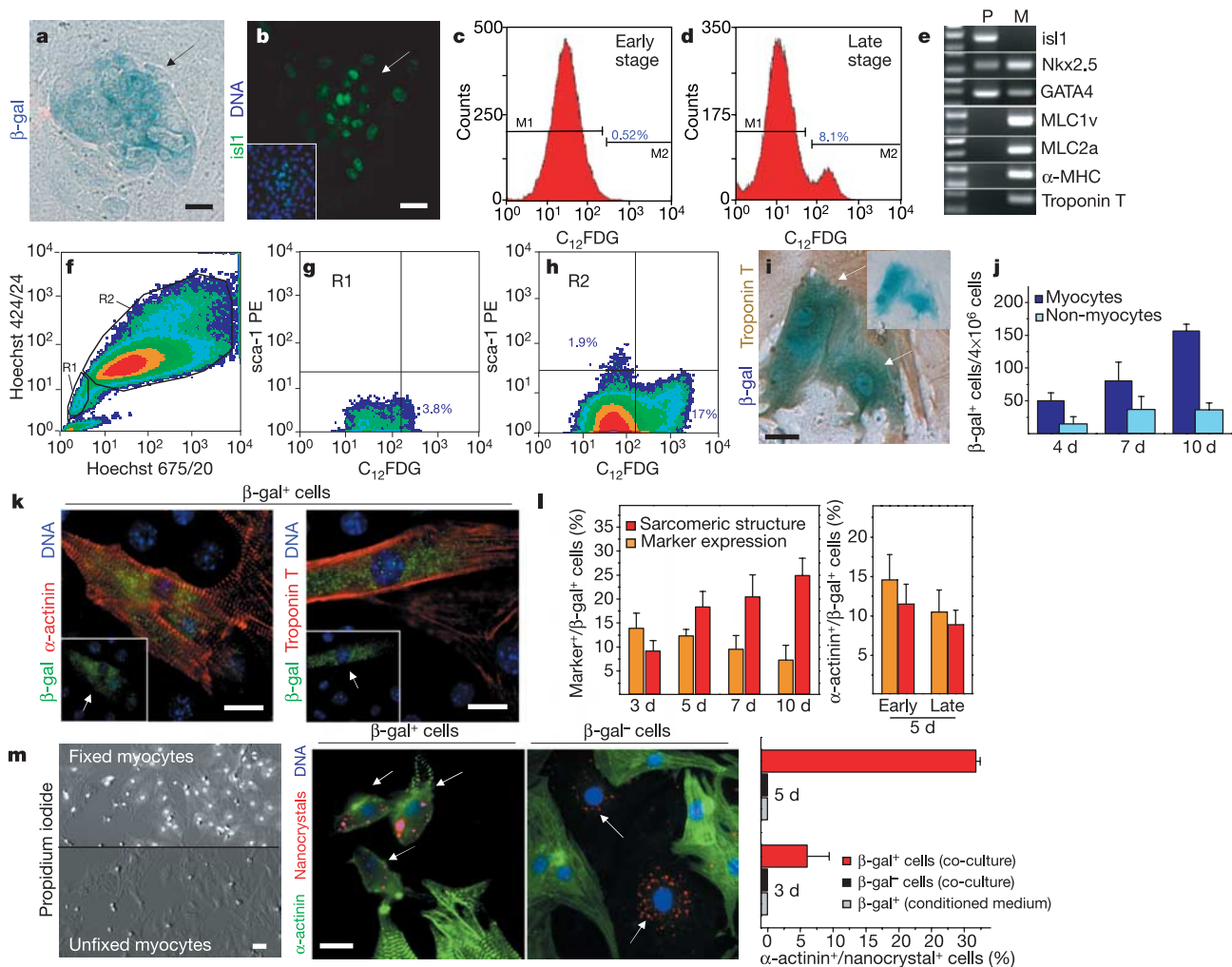


Figure 4 Amplification, characterization and myocytic differentiation of $\text{isl}1^+$ cardioblasts *in vitro*. **a, b**, Mesenchymal cell fractions from $\text{isl}1\text{-mER-Cre-mER/R26R}$ hearts at 10 days in culture after 4-OH-TM treatment. Arrows point to $\beta\text{-gal}^+$ cardioblasts detected by X-gal stain (**a**) and $\text{isl}1$ expression (**b**). Scale bars, $15 \mu\text{m}$. **c, d**, Histograms of FACS-sorted $\beta\text{-gal}^+ \text{C}_{12}\text{FDG}$ -labelled cells in cardiac mesenchymal fractions after 5 days (**c**) and 14 days (**d**) in culture. **e**, RT-PCR analysis for myocytic and progenitor markers in FACS-sorted progenitors (P) and neonatal myocytes (M). **f-h**, Flow cytometry profile of Hoechst 33342 dye efflux in cardiac mesenchymal cell fractions from double heterozygous animals after C_{12}FDG and sca-1 labelling. **i, j**, Neonatal myocytes isolated from $\text{isl}1\text{-mER-Cre-mER/R26R}$ mice at 4 days in culture. Arrows designate $\beta\text{-gal}^+$ cells coexpressing troponin T and inset shows the same cells before immunocytochemistry. Scale bar, $15 \mu\text{m}$ (**i**). Frequency of $\beta\text{-gal}^+$ myocytes and non-myocytes over time (**j**, mean value $\pm$ s.e.m., $n = 3$). Exposure to $1 \mu\text{M}$ 4-OH-TM was performed at the first day in culture.

k, Images of differentiated $\beta\text{-gal}^+$ progenitors in co-culture with wild-type neonatal myocytes. Arrows mark progenitor-derived $\beta\text{-gal}^+$ myocytes. Scale bar, $15 \mu\text{m}$. **l**, Quantification of differentiation events over time in co-culture of $\beta\text{-gal}^+$ progenitors FACS-sorted at day 10–14 (left panel) and comparison at 5 days co-culture between $\beta\text{-gal}^+$ progenitors expanded for 5 days (early) or 14 days (late) in culture (right panel). Mean values $\pm$ s.e.m. from six experiments ($n = 1,200$ cells per group). **m**, Cell fusion independent cardioblast-myocyte conversion. Propidium-iodide stain in pre-fixed myocytes (scale bar, $25 \mu\text{m}$). Arrows point to QD_{655} -labelled precursors expressing $\alpha\text{-actinin}$ in $\beta\text{-gal}^+$, but not in $\beta\text{-gal}^-$ cells during co-culture with pre-fixed myocytes. Scale bar, $15 \mu\text{m}$. The diagram represents a quantification of fusion-independent myocytic transition. Grey columns indicate the percentage of $\beta\text{-gal}^+$ cells expressing $\alpha\text{-actinin}$ in absence of pre-fixed myocytes. Mean values $\pm$ s.e.m. ($n = 3$).

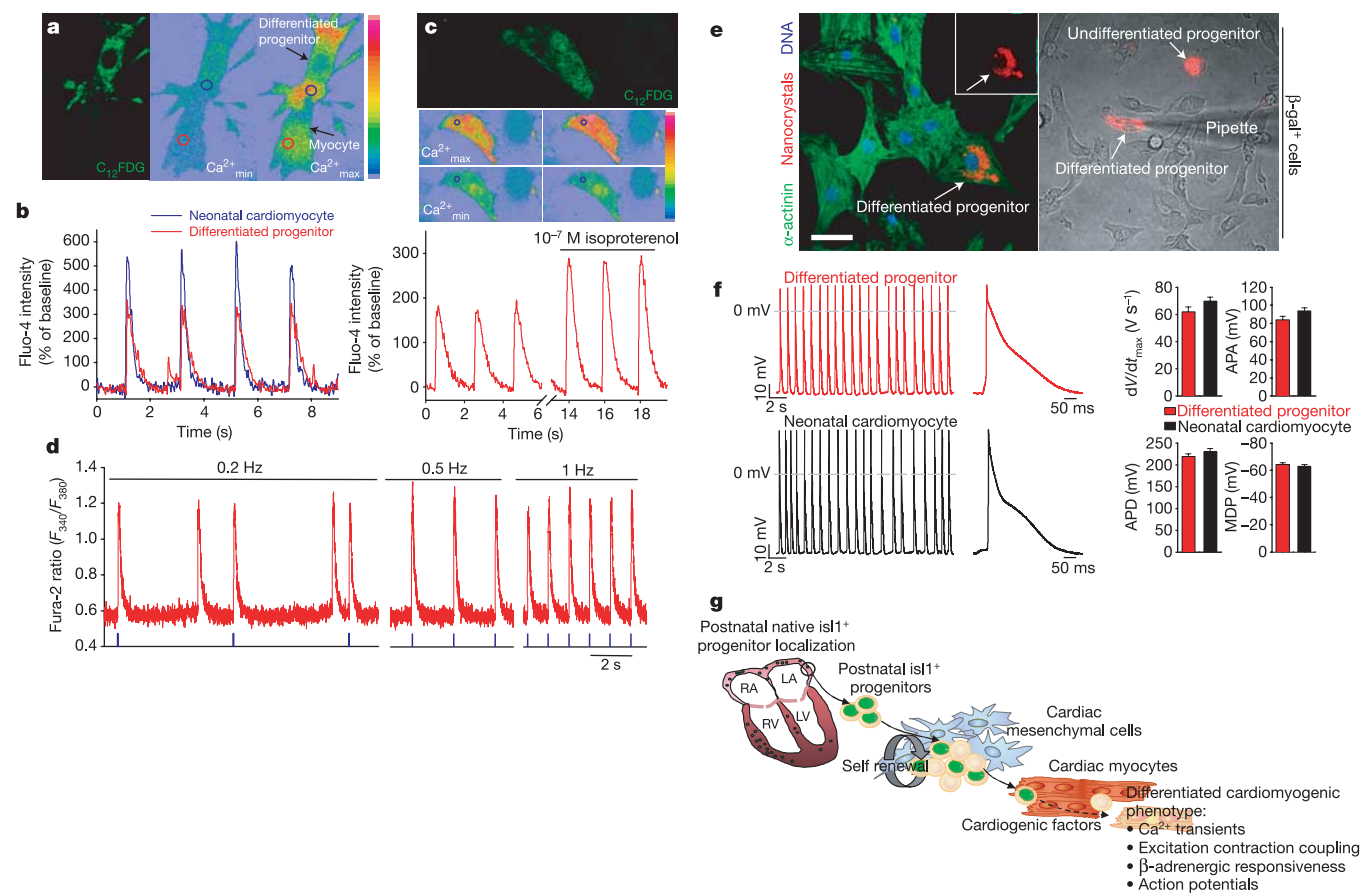


Figure 5 Real time $[Ca^{2+}]_i$ transients and action potentials in FACS-sorted β -gal $^{+}$ precursors after myocytic differentiation in co-culture. **a**, **b**, Identification of β -gal $^{+}$ cells by $C_{12}FDG$ labelling, before cell loading with the Ca^{2+} indicator fluo-4. Pseudo-colour images show minimal (Ca^{2+}_{min}) and maximal (Ca^{2+}_{max}) fluo-4 fluorescence intensity and circles indicate the Ca^{2+} measuring areas as outlined in the fluo-4 intensity traces in **b**. Eighteen measured cells: $n = 11$ differentiated progenitors, $n = 7$ neonatal myocytes. **c**, Increase in $[Ca^{2+}]_i$ transient amplitude of a β -gal $^{+}$ cell under isoproterenol (10^{-7} M) stimulation. **d**, Calcium transients of differentiated $C_{12}FDG^{+}$ precursor in response to electrical pacing at different frequencies after cell loading with the Ca^{2+} indicator fura-2.

precursors can contribute to the cardiac lineage in the absence of fusion after cellular transplantation *in vivo*, and to analyse functionally on a single-cell level the differentiated progeny²³. Cardioblasts may represent an *in vitro* system to study signalling pathways for cardiac cell lineage differentiation, which will ultimately give new insights into key steps that govern entry into specific contractile and conduction system lineages; recent work has now documented defects in later stages of ventricular lineage specification and forms of human congenital heart disease²⁴. Furthermore, this study suggests the feasibility of isolating $isl1^{+}$ cardiac progenitors from mammalian embryonic stem cell systems during cardiogenesis. □

Methods

Mice

$Isl1$ -IRES-Cre were provided by T. M. Jessel and have been previously described¹³. An IRES-Cre SV40 pA and a pgk-neomycin cassette were inserted into the exon encoding the second LIM homeodomain of $isl1$. The conditional Cre reporter mouse line R26R was generated by P. Soriano¹⁴. The $isl1$ -mER-Cre-mER targeting construct was generated by an in-frame insertion of a mER-Cre-mER SV40 pA cassette¹⁵ along with a neo-selectable

marker flanked by FRT sites into exon 1 of the genomic $isl1$ locus. The generation of $isl1$ -mER-Cre-mER knockin mice will be described in detail elsewhere. $Isl1$ -IRES-Cre/R26R and $isl1$ -mER-Cre-mER/R26R double heterozygous mice were generated by crossing single heterozygous mice. Mice are in a mixed 129 $\times$ C57BL/6 background. The $isl1$ -mER-Cre-mER line showed exclusively a tamoxifen-dependent expression of Cre. Tamoxifen (Sigma) was dissolved in corn oil and injected twice intraperitoneally in the pregnant mothers at a dose of $75 \mu g g^{-1}$ body weight at ED17 and ED18, unless specified otherwise²⁵. Animal care and all procedures were approved by the Institutional Animal Care Committee at UCSD.

Isolation of mouse cardiac progenitors and cell culture conditions

For cell isolation, we used 40–60 1–5-day-old pups that were double heterozygous for $isl1$ -mER-Cre-mER and R26R alleles. Hearts were cut into four pieces, washed repetitively in ice-cold Hank's balanced salt solution without Ca^{2+} (HBSS) and predigested overnight in $0.5 mg ml^{-1}$ trypsin-HBSS solution at $4^{\circ}C$ under constant shaking, to remove blood and dead cells. Cardiac cells were obtained by four rounds of 10 min digestions with $240 U ml^{-1}$ collagenase type II (Worthington) in HBSS at $37^{\circ}C$. The mesenchymal cell fraction containing most β -gal $^{+}$ progenitor cells was separated from myocytes and endothelial cells by two rounds of 1 h differential plating on plastic in DMEM/M199 (4:1 ratio) medium containing 10% horse serum and 5% fetal bovine serum at $37^{\circ}C$ in the presence of 5% CO_2 . Under these conditions preferentially mesenchymal and β -gal $^{+}$ precursor cells attach on plastic. After rigorous washing to remove non-adherent cells, CMC and cardioblasts were cultured for 48 h in DMEM containing penicillin ($100 U ml^{-1}$), streptomycin ($100 \mu g ml^{-1}$), HEPES (25 mM),

glutamine (2 mM), 10% newborn calf serum and 5% fetal bovine serum (FBS). Culture medium was exchanged to DMEM/F12 containing B27 supplement, 2% FBS and 10 ng ml⁻¹ EGF on the second day in culture when the CMC reached confluency, to stimulate Isl1 progenitor growth. 4-OH-TM (stock solution 1 mM in ethanol; Sigma) was applied in culture one day after cell plating at a concentration of 1 μ M and maintained for 48 h.

Methods regarding flow cytometry analysis, co-culture experiments, immunocytochemistry and lacZ staining, polymerase chain reaction with reverse transcription (RT-PCR), detection of [Ca²⁺]_i, transients and electrophysiological action potential recordings are described in the Supplementary Methods section.

Received 12 October; accepted 22 November 2004; doi:10.1038/nature03215.

- Chien, K. R. Stem cells: lost in translation. *Nature* **428**, 607–608 (2004).
- Epstein, J. A. Developing models of DiGeorge syndrome. *Trends Genet.* **17**, S13–S17 (2001).
- Chien, K. R. & Olson, E. C. Converging pathways and principles in heart development and disease: CV@CSH. *Cell* **110**, 153–162 (2002).
- Rosenthal, N. Prometheus's culture and the stem cell promise. *N. Engl. J. Med.* **346**, 267–274 (2003).
- Bischoff, R. in *Myogenesis* (eds Engel, A. G. & Franzini-Armstrong, C.) 97–118 (McGraw-Hill, New York, 1994).
- Seale, P. & Rudnicki, M. A. A new look at the origin, function, and "stem-cell" status of muscle satellite cells. *Dev. Biol.* **218**, 115–124 (2000).
- Beltrami, A. P. et al. Adult cardiac stem cells are multipotent and support myocardial regeneration. *Cell* **114**, 763–776 (2003).
- Oh, H. et al. Cardiac progenitor cells from adult myocardium: Homing, differentiation, and fusion after infarction. *Proc. Natl Acad. Sci. USA* **100**, 12313–12318 (2003).
- Balsam, L. B. et al. Haematopoietic stem cells adopt mature haematopoietic fates in ischemic myocardium. *Nature* **428**, 668–675 (2004).
- Murry, C. E. et al. Hematopoietic stem cells do not transdifferentiate into cardiac myocytes in myocardial infarcts. *Nature* **428**, 664–668 (2004).
- Nygren, J. M. et al. Bone marrow-derived hematopoietic cells generate cardiomyocytes at a low-frequency through cell fusion, but not transdifferentiation. *Nature Med.* **10**, 487–493 (2004).
- Cai, C.-L. et al. Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart. *Dev. Cell* **5**, 877–889 (2003).
- Srinivas, S. et al. Cre reporter strains produced by targeted insertion of EYFP and ECFP into the ROSA26 locus. *BMC Dev. Biol.* **1**, 4 (2001).
- Rosario, P. Generalized lacZ expression with the ROSA26 Cre reporter strain. *Nature Genet.* **21**, 70–71 (1999).
- Verrou, C., Zhang, Y., Zurn, C., Schamel, W. W. & Reth, M. Comparison of the tamoxifen regulated chimeric Cre recombinases MerCreMer and CreMer. *Biol. Chem.* **380**, 1435–1438 (1999).
- Danielian, P. S., Muccino, D., Rowitch, D. H., Michael, S. K. & McMahon, A. P. Modification of gene activity in mouse embryos *in utero* by tamoxifen-inducible form of Cre recombinase. *Curr. Biol.* **8**, 1323–1326 (1998).
- Schwenk, F., Kuhn, R., Angrand, P. O., Rajewsky, K. & Stewart, A. F. Temporally and spatially regulated somatic mutagenesis in mice. *Nucleic Acids Res.* **26**, 1427–1432 (1998).
- Martin, C. M. et al. Persistent expression of the ATP-binding cassette transporter, Abcg2, identifies cardiac SP cells in the developing and adult heart. *Dev. Biol.* **265**, 262–275 (2004).
- Wurmser, A. E. et al. Cell fusion-independent differentiation of neural stem cells to the endothelial lineage. *Nature* **430**, 350–356 (2004).
- Voura, E. B., Jaiswal, J. K., Mattoussi, H. & Simon, S. M. Tracking metastatic tumor cell extravasation with quantum dot nanocrystals and fluorescence emission-scanning microscopy. *Nature Med.* **9**, 993–998 (2004).
- Maltsev, V. A., Ji, G. J., Wobus, A. M., Fleischmann, B. K. & Hescheler, J. Establishment of β -adrenergic modulation of L-type Ca²⁺ current in the early stages of cardiomyocyte development. *Circ. Res.* **84**, 136–145 (1999).
- He, J.-Q., Ma, Y., Lee, Y., Thomson, J. A. & Kamp, T. J. Human embryonic stem cells develop into multiple types of cardiac myocytes. *Circ. Res.* **93**, 32–39 (2003).
- D'Amour, K. A. & Gage, F. H. Are somatic stem cells pluripotent or lineage-restricted? *Nature Med.* **8**, 213–214 (2002).
- Pashmforoush, M. et al. Nkx2.5 pathways and congenital heart disease: loss of ventricular myocyte lineage specification leads to progressive cardiomyopathy and complete heart block. *Cell* **117**, 373–386 (2004).
- Hayashi, S. & McMahon, A. P. Efficient recombination in diverse tissues by tamoxifen-inducible form of Cre: a tool for temporally regulated gene activation/inactivation in the mouse. *Dev. Biol.* **244**, 305–318 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank W. Giles for discussions and comments on the real-time Ca²⁺ measurements and electrophysiology, the National Center for Microscopy and Imaging Research at UCSD for the use of the high-speed multi-photon laser-scanning microscope (Bio-Rad RTS2000) and custom plug-ins for the ImageJ software (H. Hakozi, S. Chow and M. Ellisman), the UCSD Cancer Center Digital Imaging Shared Resource for deconvolution microscopy (S. McMullen and J. Feramisco). We would like especially to acknowledge D. Young for his expertise and help in the FACS sorting analysis (UCSD Cancer Center Flow Cytometry Laboratories). This work was supported by the NIH and the Jean Le Ducq Foundation. K.-L.L. is a Heisenberg Scholar of the German Research Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.R.C. (kchien@partners.org) or to S.E. (syevans@ucsd.edu).

The BRCA2 homologue Brh2 nucleates RAD51 filament formation at a dsDNA–ssDNA junction

Haijuan Yang¹, Qiubai Li¹, Jie Fan¹, William K. Holloman² & Nikola P. Pavletich^{1,3}

¹Structural Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA

²Department of Microbiology and Immunology, Hearst Microbiology Research Center, Cornell University Weill Medical College, New York, New York 10021, USA

³Howard Hughes Medical Institute, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA

The BRCA2 tumour suppressor¹ is essential for the error-free repair of double-strand breaks (DSBs) in DNA by homologous recombination^{2,3}. This is mediated by RAD51, which forms a nucleoprotein filament with the 3' overhanging single-stranded DNA (ssDNA) of the resected DSB, searches for a homologous donor sequence, and catalyses strand exchange with the donor DNA⁴. The 3,418-amino-acid BRCA2 contains eight $\sim$ 30-amino-acid BRC repeats that bind RAD51 (refs 5, 6) and a $\sim$ 700-amino-acid DBD domain that binds ssDNA⁷. The isolated BRC and DBD domains have the opposing effects of inhibiting^{8,9} and stimulating recombination⁷, respectively, and the role of BRCA2 in repair has been unclear. Here we show that a full-length BRCA2 homologue (Brh2) stimulates Rad51-mediated recombination at substoichiometric concentrations relative to Rad51. Brh2 recruits Rad51 to DNA and facilitates the nucleation of the filament, which is then elongated by the pool of free Rad51. Brh2 acts preferentially at a junction between double-stranded DNA (dsDNA) and ssDNA, with strict specificity for the 3' overhang polarity of a resected DSB. These results establish a BRCA2 function in RAD51-mediated DSB repair and explain the loss of this repair capacity in BRCA2-associated cancers.

Isolated BRC repeats can block the assembly of the RAD51 nucleoprotein filament *in vitro* when present in excess over RAD51 (refs 9, 10), and can also interfere with DSB repair in a dominant-negative fashion *in vivo*⁸. These effects have been suggested to reflect a negative control mechanism of BRCA2 over RAD51 (ref. 11). In contrast, the findings that the DBD domain (BRCA2^{DBD}) binds ssDNA and has a putative helix–turn–helix (HTH) dsDNA-binding motif⁷, coupled with the demonstration that a stoichiometric amount of the BRCA2^{DBD} can stimulate RAD51-mediated strand exchange *in vitro*⁷, indicated that BRCA2 might facilitate RAD51-mediated recombination and might have a role at the dsDNA–ssDNA (ds–ss) junction of the resected DSB⁷.

One model for reconciling these opposing biochemical activities is that BRCA2 helps to nucleate the filament by recruiting RAD51 to the DNA, and possibly to the ds–ss junction. Once there, RAD51 could dissociate from BRCA2 and partition onto ssDNA, nucleating filament formation by the pool of free RAD51. Alternatively, RAD51 could remain bound to the BRC repeat and still nucleate the filament. This is a possibility because in the RAD51–BRC crystal structure¹⁰ the BRC repeat blocks only one of the two polymerization ends of RAD51, and the second end is in principle available to bind an additional RAD51 molecule and elongate the filament in a unidirectional manner (Supplementary Fig. 1).

To test this model, we purified the full-length BRCA2 homologue Brh2 from the fungus *Ustilago maydis*. Like BRCA2, Brh2 is required for recombination, DSB repair and genomic stability¹². The 1,075-amino-acid Brh2 has homology to the DBD of human BRCA2 (ref. 13) and contains one consensus BRC repeat¹². The number of

BRC repeats is highly variable across species, with multiple BRCA2 orthologues representing a near-continuum of one to eight repeats¹⁴. As reported for the human BRCA2^{DBD} (ref. 7), we could produce soluble Brh2 only in complex with the essential Dss1 subunit¹³ (Fig. 1a and Supplementary Fig. 2). We also cloned and purified the *U. maydis* orthologue of the ssDNA-binding protein replication protein A (RPA) (Fig. 1a). RPA has a key function in RAD51-mediated recombination by removing ssDNA secondary structure that can interfere with the continuity and function of the Rad51 filament¹⁵.

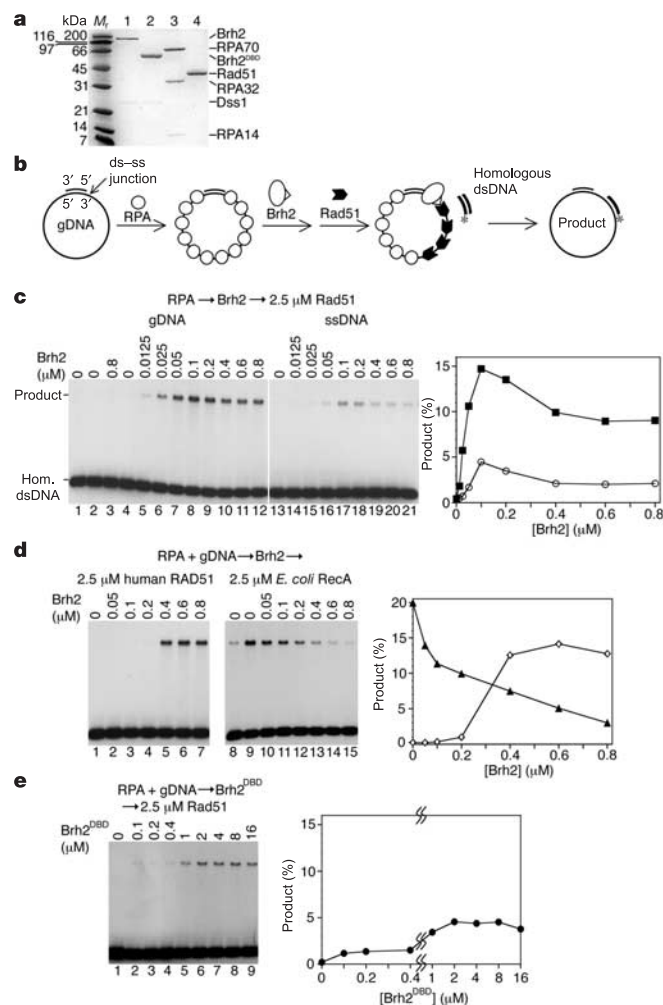


Figure 1 Substoichiometric amounts of Brh2–Dss1 stimulate Rad51-mediated strand exchange. **a**, Coomassie blue-stained SDS–polyacrylamide-gel electrophoresis of full-length Brh2–Dss1 (2 μg, lane 1), Brh2^{DBD}–Dss1 (3 μg, lane 2), heterotrimeric RPA (1.5 μg, lane 3) and Rad51 (2 μg, lane 4). *M_r*, molecular mass markers. **b**, Diagram of the strand exchange assay. The junction with the 3' overhang polarity of the resected DSB is indicated. The asterisk shows the ³²P-labelled DNA strand. dsDNA is homologous to the gDNA sequence starting 100 nt from the junction (+100 readout sequence). **c**, Autoradiogram of the strand-exchange reactions with either gDNA or ssDNA substrates. Lane 1 (no protein), lane 2 (RPA alone) and lane 3 (RPA and Brh2) are control reactions lacking Rad51. Lanes 4–21 show Brh2 titrations in the presence of gDNA or ssDNA, RPA and Rad51. The graph shows yields calculated as a percentage of the total ³²P-labelled DNA in each lane incorporated into gDNA (filled squares) or ssDNA (open circles) substrates. **d**, Brh2 stimulates human RAD51 (open diamonds) but inhibits *E. coli* RecA (filled triangles). Lane 8 has RecA but no RPA. **e**, Brh2^{DBD} stimulates Rad51 at nearly stoichiometric concentrations relative to Rad51. The Brh2 concentration axis is interrupted as indicated, to facilitate comparison with **c**.

We first investigated whether the Brh2–Dss1 complex (thereafter Brh2) can increase the probability of Rad51 nucleoprotein filament formation near a ds–ss junction. For this, we modified a strand exchange assay developed to investigate filament formation by the *Escherichia coli* RecA recombinase on ssDNA gaps¹⁶. In our assay, the ds–ss junction of a processed DSB is modelled by incorporating an 80-base-pair (bp) dsDNA region on the 5,386-nucleotide (nt) ϕ X174 (+) circular ssDNA substrate (Fig. 1b). The resulting gapped DNA (gDNA) is first incubated with RPA, then with Brh2 and last with Rad51 to form the presynaptic nucleoprotein filament. After the addition of a ³²P-labelled 80-bp homologous dsDNA, the Rad51 nucleoprotein filament catalyses strand exchange to yield a ³²P-labelled gDNA product (Fig. 1b). To measure filament occupancy specifically near the junction, we choose a dsDNA sequence homologous to the gDNA starting 100 nt from the junction (hereafter referred to as the +100 readout sequence). To be able to detect whether filament occupancy is higher near the junction than elsewhere, we use an amount of Rad51 less than the stoichiometric ratio of one Rad51 per three nucleotides¹⁷ required to cover all of the ssDNA. Under these conditions, Rad51 has been shown to form stretches of filaments distributed among the available DNA molecules instead of covering a subset of the DNA molecules entirely¹⁸.

Figure 1c shows the strand exchange reaction at the +100 readout sequence of gDNA in the presence of 2.5 μM Rad51. This corresponds to one-fifth of the stoichiometric Rad51 concentration, and was chosen as the lowest concentration with a measurable yield of 0.4% product in the absence of Brh2 (lane 4 in Fig. 1c and Supplementary Fig. 3). The addition of as little as 12.5 nM Brh2, which is only two Brh2 molecules per gDNA molecule, stimulates strand exchange, increasing the yield to 1.8%. Stimulation reaches a maximum at 100 nM Brh2, where the 14.7% yield represents a ~37-fold stimulation (Fig. 1c). The dissociation constant (*K_d*) of the Brh2–gDNA complex under these conditions is ~75 nM (Supplementary Fig. 4), consistent with the stimulation resulting from Brh2 binding to DNA. The Brh2:Rad51 ratio, which is 1:25 at the stimulation maximum and 1:200 at 12.5 nM Brh2, indicates that Brh2 acts substoichiometrically relative to Rad51. These ratios are consistent with Brh2's facilitating filament nucleation, which has been shown to be a rate-limiting step in the formation of the Rad51 filament¹⁹.

We next replaced the gDNA with the corresponding ssDNA, to which Brh2 can also bind (Supplementary Fig. 4). In the absence of Brh2, the ssDNA reaction yields 0.5% product (Fig. 1c, lane 13), essentially the same as with gDNA. Brh2 stimulates the ssDNA reaction as well, but starting at 50 nM Brh2 compared with 12.5 nM for gDNA. The stimulation maximum is still at 100 nM Brh2, but the maximum yield of 4.3% corresponds to a smaller stimulation of 8.6-fold. The 3.4-fold difference between the gDNA and ssDNA yields supports the notion that Brh2 directs filament formation to the junction. However, the ability of Brh2 to stimulate the ssDNA reaction as well is consistent with Brh2's facilitating filament nucleation. In agreement with both conclusions, Brh2 stimulates stoichiometric 12.5 μM Rad51 also, with both gDNA and ssDNA giving essentially identical yields (Supplementary Fig. 5).

These Brh2 effects are not unique to the *U. maydis* Rad51 orthologue. Human RAD51, which can bind the Brh2 BRC repeat (Supplementary Fig. 6), is also stimulated by Brh2, with a maximal yield and stimulation similar to those of *U. maydis* Rad51 (Fig. 1d). In contrast, RecA, which does not bind the Brh2 BRC repeat (Supplementary Fig. 6), is inhibited by Brh2 (Fig. 1d). As reported for human BRCA2^{DBD}, the isolated DBD domain of Brh2 (Brh2^{DBD}) can also stimulate strand exchange, giving a maximum yield of 4.5% with gDNA (Fig. 1e). However, this requires 2.0 μM Brh2^{DBD}, which corresponds to a nearly stoichiometric Brh2^{DBD}:Rad51 ratio of 0.8:1, indicating that Brh2's substoichiometric mode of action and maximal stimulation requires the coupling of the DBD and BRC domains.

We next sought to directly address whether Brh2 facilitates filament nucleation, and in particular whether it does so by displacing RPA. Although RPA is essential for recombination, it also has an inhibitory effect on Rad51 filament formation, and the requirement for accessory factors, termed mediators, to displace RPA *in vitro* is well established^{19–21}. We first examined the specificity of Brh2's effects on the RPA-supported reaction by replacing RPA with the *E. coli* ssDNA-binding (SSB) protein. As reported for other Rad51 orthologues¹⁵, SSB supports strand exchange by *U. maydis* Rad51 as well (Fig. 2a), but Brh2 cannot stimulate the SSB-supported reaction by either 2.5 μ M or 12.5 μ M Rad51.

We next investigated Brh2's effects on the kinetics of filament formation by measuring Rad51's DNA-dependent ATPase activity, used previously to demonstrate the displacement of RPA by mediator proteins such as Rad52 (refs 19–21). Figure 2b shows that the binding of 2.5 μ M Rad51 to RPA-coated gDNA gives a hydrolysis rate of 0.10 ATP per Rad51 per min, comparable to other Rad51 orthologues^{15,21}. As reported, ATP hydrolysis is preceded by a $\sim$ 7.5-min lag caused by RPA's inhibiting the slow process of filament nucleation¹⁹. Addition of 100 nM Brh2 before Rad51 addition increases the ATP hydrolysis rate 2.2-fold to 0.22, indicating that Brh2 increases the amount of Rad51 incorporated into filaments. More importantly, Brh2 also eliminates the lag phase, showing that Brh2 overcomes the inhibition of filament nucleation by RPA.

To confirm the displacement of RPA, we measured the binding of one-fifth substoichiometric Rad51 to biotin-linked gDNA that had been immobilized on streptavidin beads²². As reported²²,

preincubating the gDNA with a stoichiometric amount of RPA prevents Rad51 from binding to DNA (compare lanes 1 and 2 in Fig. 2c). The addition of Brh2 restores Rad51 filament formation while decreasing the amount of DNA-bound RPA (Fig. 2c). Even 25 nM Brh2, corresponding to about six Brh2 molecules per gDNA, substantially increases Rad51 filament formation (Fig. 2c, lane 3). Taken together with the ATPase data, these data indicate that Brh2 facilitates filament nucleation by displacing RPA at the site of nucleation.

To confirm that Brh2 nucleates the filament specifically at the

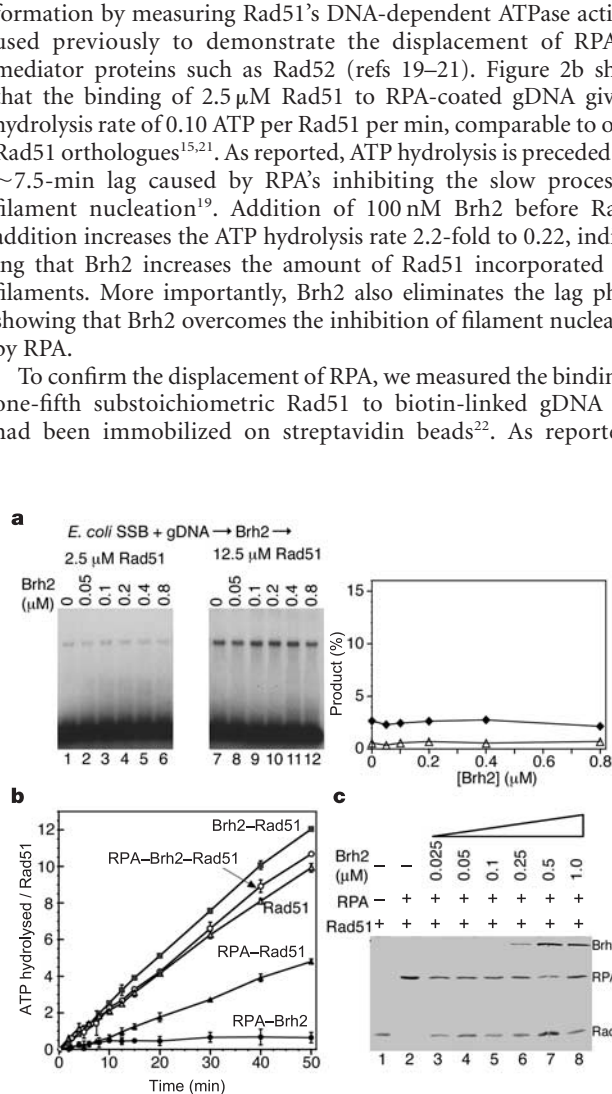


Figure 2 Brh2 facilitates filament nucleation by displacing RPA. **a**, SSB supports strand exchange mediated by both substoichiometric (lane 1) and stoichiometric (lane 7) Rad51, but neither is stimulated by Brh2. Filled diamonds, 12.5 μ M Rad51; open triangles, 2.5 μ M Rad51. **b**, Brh2 relieves the inhibition of Rad51's DNA-dependent ATPase activity by RPA. Results are means and error bars (s.d.) calculated from two independent experiments. The ATPase assay measures total DNA binding by Rad51. As reported¹⁵, Rad51's ATPase activity is higher in the absence of RPA, but the resulting filaments are incompetent in recombination presumably because they are interrupted by secondary structure. **c**, Brh2 facilitates the displacement of RPA during the formation of Rad51 nucleoprotein filaments. Immobilized gDNA was mixed with the indicated combinations of stoichiometric RPA, Brh2 and one-fifth substoichiometric Rad51. The gDNA-bound proteins were eluted with 2% SDS and analysed by SDS-polyacrylamide-gel electrophoresis and Coomassie blue staining. Only the band of the RPA70 subunit of RPA is shown.

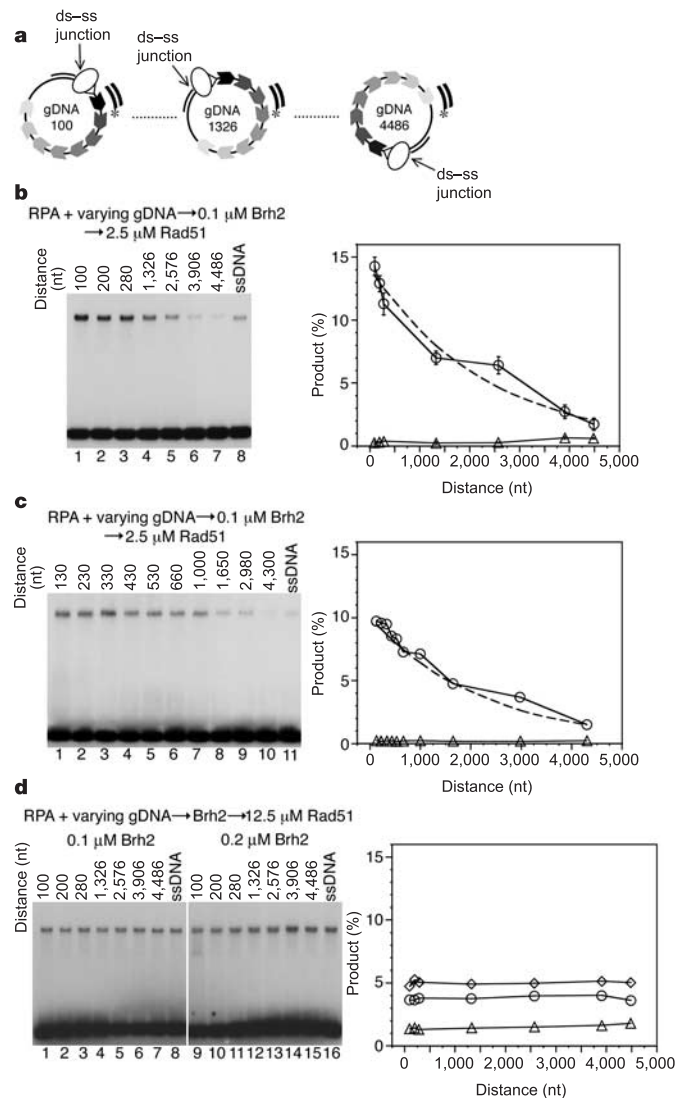


Figure 3 Strand exchange yields by substoichiometric Rad51 decrease with increasing junction-readout sequence distance. **a**, Diagram of varying junction-readout sequence distance by moving the double-stranded region of gDNA relative to a constant readout sequence, shown approximately for the 100-nt, 1,326-nt and 4,486-nt distances. **b**, Strand exchange yields at the +100 readout sequence with the set of gDNAs and ssDNA described in Methods. The graph shows yields with 100 nM Brh2 (circles; averages and error bars (s.d.) from three independent experiments), without Brh2 (triangles; 0.4% average, 0.1% standard deviation (s.d.) for the seven distances), and the fitted exponential curve $y = 14.1e^{-(1/2,325)x}$ (dashed line); $R^2 = 0.96$. **c**, Strand exchange yields with a second readout sequence and set of gDNAs and ssDNA described in Methods. The symbols in the graph are as in **b**; the equation for the exponential curve is $y = 10.4e^{-(1/2,320)x}$, $R^2 = 0.98$. **d**, Strand exchange with stoichiometric Rad51 performed as in **b** shows no distance dependence with 0 μ M Brh2 (triangles; 1.5% average, 0.2% s.d.), 0.1 μ M Brh2 (circles; 3.8% average, 0.2% s.d.) or 0.2 μ M Brh2 (diamonds; 4.9% average, 0.3% s.d.).

ds-ss junction, we sought to compare strand exchange yields at increasing junction-readout sequence distances. We reasoned that if Brh2 causes the filament to nucleate at the junction of gDNA, then the use of one-fifth substoichiometric Rad51 should result in a distribution of filament lengths that on average extend to one-fifth the length of the gDNA from the junction. Then, the probability of a readout sequence's being covered by a filament should decrease with increasing distance from the junction, resulting in a corresponding decrease in the strand exchange yield. Because the base composition of the recombining DNA can affect strand exchange efficiency²³, we varied the distance by moving the double-stranded region on gDNA and keeping the +100 readout sequence the same (Fig. 3a).

We measured strand exchange yields by 2.5 μ M Rad51 and 100 nM Brh2 at seven junction-readout sequence distances ranging from 100 to 4,486 nt. Figure 3b shows that the yields decrease 7.4-fold over this range, confirming a key prediction of the targeting model. The yields decrease continuously, starting with the shortest distance. This shows that filament occupancy is highest closest to the junction, indicating that most of the filaments nucleate at the junction. That there is little junction-independent nucleation is also evident because the yields at the two farthest distances are lower than the Brh2-stimulated yield with ssDNA (Fig. 3b, lanes 6, 7 and 8). However, the yields at the farthest distances are still higher than in the absence of Brh2, indicating that a significant number of junction-nucleated filaments extend several thousand nucleotides. The strand exchange yields are essentially constant across the seven distances in the absence of Brh2 (Fig. 3b).

The data also show that Brh2 does not stimulate filament formation at the junction having 5' overhang ssDNA, revealing a strict specificity for the junction with the polarity of a resected DSB²⁴ (Fig. 3b). Although part of this specificity is probably due to Brh2's ~2.5-fold higher affinity for 3' overhang DNA compared with 5' overhang or ssDNA (Supplementary Fig. 7), it could also reflect the filament's being nucleated by a BRC-bound Rad51, which would result in unidirectional elongation.

The decrease in yields with increasing distance is best fitted with an exponential function $y = 14.1e^{-(1/2,325)x}$ ($R^2 = 0.96$; Fig. 3b), where x is the distance and y is the recombination yield (or Product), and the exponential constant 1/2,325 describes how yields decrease with distance (Supplementary Fig. 8). An exponential decay would be predicted by our junction-nucleation model if we assume that the probability of filament elongation at each nucleotide is constant along the length of the gDNA.

To confirm that the distance dependence is independent of the choice of readout sequence and set of gDNAs, we tested a different readout sequence with a set of gDNA molecules sampling ten distances, ranging from 130 to 4,300. The results show a distance dependence very similar to that of the first readout sequence, with an exponential constant of 1/2,320 ($R^2 = 0.98$; Fig. 3c). We also tested the model's prediction that there should be no distance dependence with stoichiometric Rad51. Figure 3d shows that in the presence of either 100 nM Brh2 or 200 nM Brh2, which is the stimulation maximum for 12.5 μ M Rad51 (Supplementary Fig. 5), the yields are constant across the seven distances.

Taken together, our results show that Brh2 facilitates filament nucleation by recruiting Rad51 to DNA and displacing RPA at the nucleation site. In the presence of a junction, most of the filaments that Brh2 nucleates start at the junction, with a high level of specificity for the junction with the polarity of a processed DSB.

There are parallels between Brh2 and the mediator protein Rad52 (refs 19–21,25) in that they both can facilitate Rad51 nucleoprotein filament formation by displacing RPA. Unlike Brh2, however, Rad52 seems to act at higher concentrations, possibly stoichiometrically to RPA, with which it interacts^{20,21,26}, and it is not yet clear whether Rad52 has a preferred site of action on the resected DSB. There are also several similarities between Brh2 and the *E. coli* RecFOR

complex, which facilitates RecA filament formation specifically on ssDNA gaps¹⁶. RecFOR also recognizes ds-ss junctions, acts at substoichiometric concentrations relative to RecA, and accelerates filament formation by displacing SSB¹⁶.

The initiation of homology-directed DSB-repair is a highly regulated process in which RPA is central. RPA localizes to DSBs early on in the process, well before Rad51 does, and coats the resected ssDNA^{27,28}. By sequestering the ssDNA and making Rad51-filament formation dependent on accessory factors, RPA might also be central in coordinating and regulating this process. In *Saccharomyces cerevisiae*, which does not have a recognizable BRCA2 homologue, Rad52 is required for Rad51-filament formation in DSB repair^{22,25} and in other Rad51-mediated processes²⁹. This is an essential function because *rad52* mutations cause hypersensitivity to DNA-damaging agents and defects in other Rad51-mediated recombination processes²⁹. Unlike in *S. cerevisiae*, in mice the deletion of *rad52* does not cause sensitivity to DSB-inducing agents³⁰, although it does affect other recombination processes. In this respect, *U. maydis* is more similar to mammalian cells, because *rad52* deletion does not affect DSB repair (W.K.H., unpublished data). The results presented here indicate that in higher eukaryotes BRCA2 evolved as the factor that nucleates the Rad51 filament for the specialized case of DSB repair. □

Methods

Protein expression and purification

The overexpression of soluble Brh2 or Brh2^{DBD} requires co-expression with the *U. maydis* Dss1 subunit¹³, so all experiments used purified Brh2–Dss1 complex. The complex was produced by infecting Hi5 insect cells with a pFastBac DUAL (Invitrogen) baculovirus expressing glutathione S-transferase (GST)-tagged full-length Brh2 and untagged Dss1, and was purified by glutathione affinity chromatography, cleavage of the GST tag with tobacco etch virus protease (TEV), cation exchange and anion exchange, and gel-filtration chromatography¹³. The boundaries of Brh2^{DBD} (residues 539–1,075) were determined by limited proteolysis, and the Brh2^{DBD}–Dss1 protein complex was produced and purified similarly.

The *U. maydis* RPA open reading frames were amplified from a complementary DNA library and sequenced. The heterotrimeric RPA complex was produced by co-infecting Hi5 insect cells with a pFastBac1 baculovirus expressing the untagged 70-kDa subunit and a pFastBac DUAL baculovirus expressing the GST-tagged 32-kDa and untagged 14-kDa subunits. RPA was purified by glutathione affinity chromatography, TEV cleavage, anion exchange, and gel-filtration chromatography. The Brh2–Dss1, Brh2^{DBD}–Dss1 and RPA complexes were stored in 20 mM Tris-HCl pH 7.6, 200 mM NaCl, 10 mM dithiothreitol (DTT), 10% glycerol at –80 °C.

His₆-tagged *U. maydis* Rad51 was overexpressed in the *E. coli* BL21 (DE3) Δ recA strain from a modified pET15b vector, and was purified by ammonium sulphate fractionation, Ni²⁺ affinity chromatography, thrombin cleavage of the His₆ tag, and cation-exchange and anion-exchange chromatography. Purified Rad51 was dialysed and stored in 20 mM KH₂PO₄ pH 7.4, 100 mM KCl, 10 mM DTT, 10% glycerol at –80 °C.

DNA substrates

Single-stranded ϕ X174 viral (+) strand DNA (ss ϕ X174) was from New England Biolabs. All oligonucleotides were synthesized and gel purified by Sigma Genosys. In Figs 1 and 2 the gDNA was prepared by annealing ss ϕ X174 to an 80-nt oligonucleotide (complementary to nt 5,207–5,286 of ss ϕ X174), and was purified with CHROMA SPIN+TE-400 columns (BD Biosciences). The ssDNA was mock-annealed with a non-complementary 80-nt oligonucleotide of the same sequence as nt 5,207–5,286 of ss ϕ X174. The blunt-ended, 80-bp dsDNA substrate homologous to the +100 readout sequence was ³²P-labelled on the strand complementary to nt 1–80 of ss ϕ X174. The gDNA substrates in Fig. 3 are described in Supplementary Fig. 8.

Strand exchange assay

Buffer R, containing 35 mM MOPS pH 7.2, 2.5 mM ATP, 3 mM magnesium acetate, 25 mM NaCl, 1 mM DTT, 2.5 μ g ml^{–1} BSA, 12 mM creatine phosphate and 30 μ g ml^{–1} creatine kinase was used for all reactions. All incubation steps were at 37 °C. gDNA or ssDNA (6.5 nM molecules corresponding to 37.5 μ M nucleotides) was first incubated with 2 μ M RPA in 7.4 μ l for 5 min. Brh2, or just protein storage buffer where applicable, was added in 0.6 μ l, followed by incubation for a further 5 min, after which Rad51 was added in 4 μ l of buffer R. After 5 min, ³²P-labelled homologous dsDNA substrate (5 nM molecules) in 1 μ l was added to start the reaction. All protein concentrations in the text refer to the final concentrations. After 2 h, reactions were stopped by treatment for 30 min with proteinase K (1 mg ml^{–1}) and SDS (0.5%) at 37 °C, and were separated by 1.1% agarose gel electrophoresis in Tris-borate-EDTA buffer containing 5 μ g ml^{–1} ethidium bromide. The dried gels were exposed to PhosphorImager (Fuji) plates, and were scanned with Fujifilm FLA-5000. The images were quantified with ImageGauge Software. The percentage of product was calculated as radiolabelled gDNA product divided by the total amount of radioactive DNA in each lane.

ATPase assay

The ATPase assay was assembled as in the strand exchange assay, except that buffer R contained 1 mM ATP supplemented with [γ - 32 P]ATP (10 μ Ci per reaction), and lacked creatine phosphate and creatine kinase. gDNA (37.5 μ M nucleotides) was first incubated with RPA (2 μ M) for 5 min and then with Brh2 (0.1 μ M) for a further 5 min. The reactions were started by adding 2.5 μ M Rad51. Where a protein was omitted, storage buffer was added instead. At the indicated times, 2 μ l of each reaction was mixed with 0.7 μ l of 5 M formic acid and spotted on a polyethyleneimine cellulose (PEI) paper (EM Sciences). The PEI paper was developed in 1 M formic acid and 0.5 M LiCl, dried, and quantified by phosphorimaging.

Biotin pull-down assay

A 5'-biotinylated oligonucleotide (complementary to nt 1–80 of ϕ X174) was annealed to ϕ X174 (+) and coupled to magnetic streptavidin beads as described²². A 10 μ l suspension of beads, containing ~80 ng of gDNA, was first incubated with stoichiometric RPA (1.3 μ M), then with the indicated amounts of Brh2, and finally with one-fifth substoichiometric Rad51 (1.6 μ M) in buffer R supplemented with 0.01% Igepal. The beads were washed, and bound proteins were eluted as described²². Background protein binding was typically ~5%.

Received 24 August; accepted 1 December 2004; doi:10.1038/nature03234.

1. Rahman, N. & Stratton, M. R. The genetics of breast cancer susceptibility. *Annu. Rev. Genet.* **32**, 95–121 (1998).
2. Moynahan, M. E., Pierce, A. J. & Jasin, M. BRCA2 is required for homology-directed repair of chromosomal breaks. *Mol. Cell* **7**, 263–272 (2001).
3. Tutt, A. *et al.* Mutation in *Brca2* stimulates error-prone homology-directed repair of DNA double-strand breaks occurring between repeated sequences. *EMBO J.* **20**, 4704–4716 (2001).
4. Sung, P. Catalysis of ATP-dependent homologous DNA pairing and strand exchange by yeast RAD51 protein. *Science* **265**, 1241–1243 (1994).
5. Wong, A. K., Pero, R., Ormonde, P. A., Tavtigian, S. V. & Bartel, P. L. RAD51 interacts with the evolutionarily conserved BRC motifs in the human breast cancer susceptibility gene *brca2*. *J. Biol. Chem.* **272**, 31941–31944 (1997).
6. Chen, P. L. *et al.* The BRC repeats in BRCA2 are critical for RAD51 binding and resistance to methyl methanesulfonate treatment. *Proc. Natl Acad. Sci. USA* **95**, 5287–5292 (1998).
7. Yang, H. *et al.* BRCA2 function in DNA binding and recombination from a BRCA2–DSS1–ssDNA structure. *Science* **297**, 1837–1848 (2002).
8. Chen, C. F., Chen, P. L., Zhong, Q., Sharp, Z. D. & Lee, W. H. Expression of BRC repeats in breast cancer cells disrupts the BRCA2–Rad51 complex and leads to radiation hypersensitivity and loss of G₂/M checkpoint control. *J. Biol. Chem.* **274**, 32931–32935 (1999).
9. Davies, A. A. *et al.* Role of BRCA2 in control of the RAD51 recombination and DNA repair protein. *Mol. Cell* **7**, 273–282 (2001).
10. Pellegrini, L. *et al.* Insights into DNA recombination from the structure of a RAD51–BRCA2 complex. *Nature* **420**, 287–293 (2002).
11. West, S. C. Molecular views of recombination proteins and their control. *Nature Rev. Mol. Cell Biol.* **4**, 435–445 (2003).
12. Kojic, M., Kostub, C. F., Buchman, A. R. & Holloman, W. K. BRCA2 homolog required for proficiency in DNA repair, recombination, and genome stability in *Ustilago maydis*. *Mol. Cell* **10**, 683–691 (2002).
13. Kojic, M., Yang, H., Kostub, C. F., Pavletich, N. P. & Holloman, W. K. The BRCA2-interacting protein DSS1 is vital for DNA repair, recombination, and genome stability in *Ustilago maydis*. *Mol. Cell* **12**, 1043–1049 (2003).
14. Lo, T., Pellegrini, L., Venkataraman, A. R. & Blundell, T. L. Sequence fingerprints in BRCA2 and RAD51: implications for DNA repair and cancer. *DNA Repair* **2**, 1015–1028 (2003).
15. Sugiyama, T., Zaitseva, E. M. & Kowalczykowski, S. C. A single-stranded DNA-binding protein is needed for efficient presynaptic complex formation by the *Saccharomyces cerevisiae* Rad51 protein. *J. Biol. Chem.* **272**, 7940–7945 (1997).
16. Morimatsu, K. & Kowalczykowski, S. C. RecFOR proteins load RecA protein onto gapped DNA to accelerate DNA strand exchange: a universal step of recombinational repair. *Mol. Cell* **11**, 1337–1347 (2003).
17. Sung, P. & Roberson, D. L. DNA strand exchange mediated by a RAD51–ssDNA nucleoprotein filament with polarity opposite to that of RecA. *Cell* **82**, 453–461 (1995).
18. McIlwraith, M. J. *et al.* Reconstitution of the strand invasion step of double-strand break repair using human Rad51 Rad52 and RPA proteins. *J. Mol. Biol.* **304**, 151–164 (2000).
19. Sugiyama, T. & Kowalczykowski, S. C. Rad52 protein associates with replication protein A (RPA)-single-stranded DNA to accelerate RAD51-mediated displacement of RPA and presynaptic complex formation. *J. Biol. Chem.* **277**, 31663–31672 (2002).
20. New, J. H., Sugiyama, T., Zaitseva, E. & Kowalczykowski, S. C. Rad52 protein stimulates DNA strand exchange by Rad51 and replication protein A. *Nature* **391**, 407–410 (1998).
21. Shinohara, A. & Ogawa, T. Stimulation by Rad52 of yeast Rad51-mediated recombination. *Nature* **391**, 404–407 (1998).
22. Wolner, B., van Komen, S., Sung, P. & Peterson, C. L. Recruitment of the recombinational repair machinery to a DNA double-strand break in yeast. *Mol. Cell* **12**, 221–232 (2003).
23. Gupta, R. C., Folta-Stogniew, E., O'Malley, S., Takahashi, M. & Radding, C. M. Rapid exchange of A:T base pairs is essential for recognition of DNA homology by human Rad51 recombination protein. *Mol. Cell* **4**, 705–714 (1999).
24. White, C. I. & Haber, J. E. Intermediates of recombination during mating type switching in *Saccharomyces cerevisiae*. *EMBO J.* **9**, 663–673 (1990).
25. Sugawara, N., Wang, X. & Haber, J. E. *In vivo* roles of Rad52, Rad54, and Rad55 proteins in Rad51-mediated recombination. *Mol. Cell* **12**, 209–219 (2003).
26. Sung, P. Function of yeast Rad52 protein as a mediator between replication protein A and the Rad51 recombinase. *J. Biol. Chem.* **272**, 28194–28197 (1997).
27. Wang, X. & Haber, J. E. Role of *Saccharomyces* single-stranded DNA-binding protein RPA in the strand invasion step of double-strand break repair. *PLoS Biol.* **2**, 0104–0112 (2004).
28. Lisby, M. L., Barlow, J. H., Burgess, R. C. & Rothstein, R. Choreography of the DNA damage response:

- spatiotemporal relationships among checkpoint and repair proteins. *Cell* **118**, 699–713 (2004).
29. Petes, T. D., Malone, R. E. & Symington, L. S. in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Genome Dynamics, Protein Synthesis, and Energetics* (eds Broach, J. R., Jones, E. & Pringle, J.) 407–521 (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1991).
 30. Rijkers, T. *et al.* Targeted inactivation of mouse RAD52 reduces homologous recombination but not resistance to ionizing radiation. *Mol. Cell. Biol.* **18**, 6423–6429 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank H. Erdument-Bromage of the Sloan-Kettering Microchemistry Facility and D. King from the HHMI mass spectrometry laboratory at U.C. Berkeley for N-terminal sequence and mass spectroscopic analysis; and P. D. Jeffrey and members of the Pavletich laboratory for helpful discussions. This work was supported by the NIH, the Howard Hughes Medical Institute, and the Arthur and Rochelle Belfer Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.P.P. (nikola@xray2.mskcc.org).

Structural basis for substrate binding, cleavage and allostery in the tRNA maturase RNase Z

Inés Li de la Sierra-Gallay², Olivier Pellegrini¹ & Ciarán Condon¹

¹CNRS UPR 9073, affiliated with Université de Paris 7—Denis Diderot, ²FRC550, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005 Paris, France

Transfer RNAs (tRNAs) are synthesized as part of longer primary transcripts that require processing of both their 3' and 5' extremities in every living organism known. The 5' side is processed (matured) by the ubiquitously conserved endonucleolytic ribozyme, RNase P¹, whereas removal of the 3' tails can be either exonucleolytic^{2,3} or endonucleolytic⁴. The endonucleolytic pathway is catalysed by an enzyme known as RNase Z, or 3' tRNase^{5,6}. RNase Z cleaves precursor tRNAs immediately after the discriminator base (the unpaired nucleotide 3' to the last base pair of the acceptor stem, used as an identity determinant by many aminoacyl-tRNA synthetases) in most cases^{6–8}, yielding a tRNA primed for addition of the CCA motif by nucleotidyl transferase. Here we report the crystal structure of *Bacillus subtilis* RNase Z at 2.1 Å resolution, and propose a mechanism for tRNA recognition and cleavage. The structure explains the allosteric properties of the enzyme, and also sheds light on the mechanisms of inhibition by the CCA motif and long 5' extensions. Finally, it highlights the extraordinary adaptability of the metallo-hydrolase domain of the β -lactamase family for the hydrolysis of covalent bonds.

RNase Z is found in the vast majority, if not all, eukaryotes and archaea, and in about half of the sequenced bacteria. It is often essential for growth⁹, and mutations in one of the two genes encoding RNase Z (*ELAC2*) have been linked with prostate cancer in man¹⁰. The products of the cleavage reaction are a 3'-matured tRNA ending in the hydroxyl group of the discriminator base and a trailer sequence with a 5' phosphate⁷. A variant of this enzyme, which can cleave immediately following the CCA motif, thus producing mature tRNA in a single step, was recently discovered in *Thermatoga maritima*¹¹. In many organisms, RNase Z activity is severely inhibited by the presence of a CCA motif in the pre-tRNA^{6,9,12}, thus preventing futile cycles of CCA removal and re-addition. RNase Z activity is also inhibited by long 5' extensions in bacteria, plants and animals, suggesting that for many tRNAs,

RNase P processing of the 5' end precedes cleavage by RNase Z *in vivo*^{7,9}.

Sequence homology searches predict that RNase Z belongs to the β -lactamase family of the zinc-dependent metallo-hydrolases (Supplementary Fig. 1), and a mutational analysis of residues equivalent to those responsible for Zn coordination in β -lactamase has confirmed their importance for RNase Z activity¹¹. We solved the crystal structure of seleno-methionine-labelled RNase Z by multiple anomalous dispersion (MAD). The crystallographic and refinement data are provided in Supplementary Table 1. We find one homodimer of RNase Z per asymmetric unit, and the enzyme indeed bears significant similarity to the Zn-dependent metallo- β -lactamases. The core domain of each monomer consists of two opposing $\beta\beta\beta\alpha\beta\alpha\beta$ motifs that form two seven-strand β -sheets sandwiched on either side by α -helices. The first four β -strands ($\beta 7$, $\beta 8$, $\beta 9$, $\beta 12$ and $\beta 16$, $\beta 1$, $\beta 2$, $\beta 3$) are antiparallel and the next three are parallel in each of the β -sheets (Fig. 1 and Supplementary Fig. 2). The structure is characterized by a long flexible arm (showing no similarity to proteins of known structure) extending from the core β -lactamase domain between β -strands 9 and 12. The dimerization interface is formed by helices $\alpha 1$, $\alpha 2$, $\alpha 3$, and the $\beta 1$ - $\beta 2$ loop of each monomer. The main hydrogen-bond contacts (<3.3 Å) are between residues Ala 41, Gln 43, His 44, Ser 103 and Thr 107 of one monomer and Asn 18, Glu 40 and Tyr 70 of the other. Reciprocal inter-subunit hydrogen bonds are formed between Asn 18–His 44, Glu 40–Gln 43, Glu 40–Ala 41, and between Tyr 70 and both Thr 107 and Ser 103.

The monomers are clearly not equivalent in the crystallographic structure (Fig. 1). In addition to the long flexible arm, visible in the B subunit, but not in A, we could also resolve helix $\alpha 7$, and two Zn atoms and a phosphate ion in the catalytic site of the A subunit that were absent from its partner. Furthermore, residues 65–69 adopt a 3/10 helical structure in A, whereas these residues are α -helical in the B subunit. The protruding arm is the most distinguishing feature between RNase Z and other members of the Zn-dependent metallo- β -lactamase family (Fig. 2), and we believe that it plays a role in substrate recognition (see below). Superposition of the A and

B subunits shows that, whereas the α -carbons of those regions of RNase Z near the dimerization interface (amino acids 1–136) can be superimposed with a root mean square deviation (r.m.s.d.) as low as 0.27 Å, those visible in the rest of the enzyme, including the predicted active site, have an r.m.s.d. of 2.51 Å, emphasizing the structural differences in this part of the protein.

The phosphate ion fortuitously co-crystallized in the catalytic site of the A subunit is held in place by potential hydrogen bonds with water molecules 5 and 147, histidines at positions 140, 247 and 269, aspartates 211 and 67, and coordination by the two Zn^{2+} ions (see below). We manually docked yeast tRNA^{Phe} (PDB code: IEHZ) on RNase Z (Fig. 3a and b) by superimposing the phosphate group of the scissile phosphodiester bond on the phosphate ion in the active site and performing an energy minimization using CNS¹³. Because the CCA motif is known to inhibit RNase Z activity, the cytidines at positions 74 and 75, downstream of the cleavage site, were changed to adenine and the orientation of nucleotides (nts) 74–76 twisted slightly to allow them to exit the protein rather than getting buried in it. Only minor movements of the loops (<2 Å) around the contact points were required to accommodate the tRNA. By leaning the top of the flexible arm forward by about 6 Å, it makes contacts with the T-arm of the tRNA, 'clamping' it in position. Interestingly, the regions of the protein that had to be moved to accommodate the tRNA overlapped greatly with those that demonstrated flexibility when the two subunits were superimposed. The tRNA lies across the top of the B subunit, with the acceptor stem making contacts with positively charged residues in the loops between $\beta 2$ and $\beta 3$ (Arg 31), and $\alpha 1$ and $\beta 4$ (Lys 52, Arg 54, Lys 55). The charged amino acids in these loops and in the flexible arm of the B subunit (lysines 164, 176, 178, 179) can be readily seen in the surface plot of electrostatic

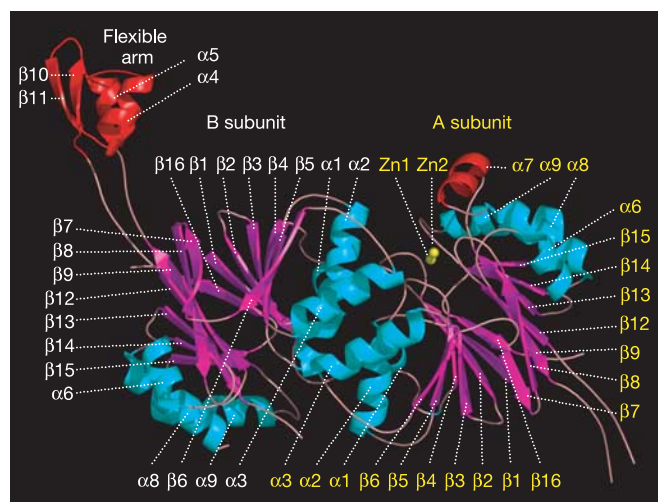


Figure 1 Ribbon representation of RNase Z at 2.1 Å resolution. Features of the A subunit are labelled in yellow, those of the B subunit in white. α -helices are shown in turquoise; β -sheets in mauve. Structural elements visible in only one subunit are shown in red. Zn^{2+} ions are represented as yellow spheres. All 307 amino acids are represented in one or other of the subunits. The flexible arm visible in the B subunit is indicated. RNase Z used in this study had a C-terminal histidine tag and two mutations, Ile46Met and Leu228Met, to increase the total number of methionine residues to five per monomer for seleno-methionine labelling. Seleno-methionine-labelled RNase Z has wild-type activity levels *in vitro* (data not shown).

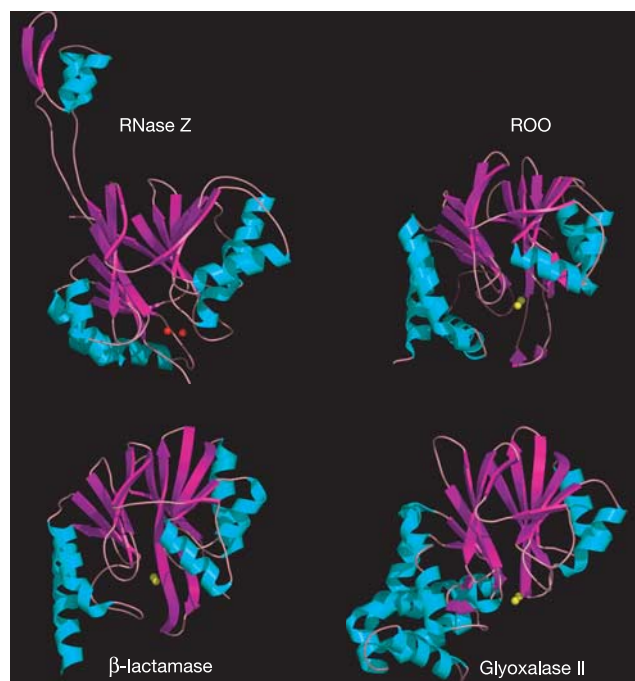


Figure 2 Structural comparison of members of the metallo- β -lactamase family. The monomer structures of RNase Z, rubredoxin:oxygen oxidoreductase (ROO; PDB²⁹ ID code 1E5D), β -lactamase IMP-1 (1JJE) and glyoxalase II (1QH5) are shown. The B subunit of RNase Z is shown with water molecules (red spheres) in the pseudo-active site. ROO contains iron in the active site and its flavodoxin-like domain has been removed for simplicity. All ions are represented as yellow spheres. Values of r.m.s.d. are 1.92 Å for ROO (107 superimposed α -carbons), 2.31 Å for β -lactamase (126 superimposed α -carbons) and 1.73 Å for glyoxalase (116 superimposed α -carbons) compared to the B subunit of RNase Z.

potential shown in Fig. 3c. Thus, the tRNA is recognized primarily by one subunit but cleaved by the other. The fact that the only contacts between RNase Z and the tRNA are with the T-arm and acceptor stem nicely explains the results of previous studies showing that these two domains are sufficient for cleavage *in vitro*¹⁴.

On the basis of the cleavage mechanisms of other metallo-hydrolases of this family, we can propose a mechanism by which RNase Z cleaves precursor tRNA. The two catalytic Zn^{2+} ions of RNase Z (Fig. 4a) are co-ordinated in a manner very similar to that

seen in other metallo-hydrolases, particularly that of human glyoxalase II¹⁵. Five histidines, two aspartate residues, a water molecule and the phosphate ion provide a potential six-way, pseudo-octahedral coordination of the Zn^{2+} ions. Asp 67 acts as a general base to remove the proton from water molecule 147, allowing the hydroxide ion to perform a nucleophilic attack on the phosphate group polarized by the two Zn^{2+} ions (Fig. 4b). This results in cleavage of the phosphodiester linkage between nt 73 and nt 74 of the tRNA precursor. The 3' oxygen of the discriminator base (nt 73) is then stabilized by taking the proton from the nearby water 5, resulting in a tRNA molecule with a 3' OH group.

His 68 is stabilized by hydrogen bonds with Asp 37 and Thr 62. Mutations in Asp 37 inactivate RNase Z (ref. 11, and our unpublished data). Thus, Asp 37 clearly plays a key role in maintaining the catalytic site in an active conformation. Interestingly, Asp 37 is the only amino acid in a disallowed region in the Ramachandran plot, although its electron density is clearly visible in both subunits of the structure (see, for example, Fig. 4a).

RNase Z activity is inhibited both by long 5' extensions (>10 nt) and by the presence of a CCA motif downstream of the discriminator base. Inhibition by 5' extensions can be explained by a narrowing of the channel that accommodates the tRNA just at the point where the acceptor arm becomes single stranded (Fig. 3b, inset, and c). In our model, the 5' guanosine of the tRNA butts up against the loop between helix $\alpha 3$ and strand $\beta 6$, and only the single stranded extension continues towards the active site. We believe that only short 5' extensions are tolerated because of the statistical probability that further base pairing occurs with longer extensions, and that the channel is only wide enough for one strand. Consistent with this idea, RNase Z activity is completely inhibited by 5' extensions of only 3 nt that are perfectly paired with the beginning of the 3' trailer¹⁶.

The CCA motif is also an antideterminant for RNase Z activity¹². Although the first C residue of CCA (C74) accounts for most of the inhibition of the *B. subtilis* enzyme, the second cytosine causes additional inhibition in the double-C context⁹. The primary difference between cytidine, which is inhibitory, and uridine, for example, which is accepted in position 74, is the replacement of the oxygen at position C4 by an amino group in cytidine. This, coupled with the particular stacking conformation of the cytidines on the tRNA acceptor stem, results in a clash between the amino group of C74 and the loop between strands $\beta 1$ and $\beta 2$ that we predict does not occur with other bases. The highly conserved Pro 13 is likely to play an important role either in direct discrimination or by rigidifying the loop and thus reducing the tolerance for other bases. Interestingly, the *T. maritima* enzyme, which does not discriminate against the CCA motif, has an N-terminal truncation that removes the first 14 residues relative to *B. subtilis* RNase Z. Two glutamine residues (Gln 43 and 45), proposed to be part of a CCA binding site in *T. maritima* RNase Z¹¹, are located in helix $\alpha 1$ and seem too far from the active site to play a direct role in CCA recognition.

Previous studies had suggested that *Escherichia coli* RNase Z (ElaC) had allosteric Zn-dependent phosphodiesterase activity (ref. 17 and our unpublished data). It is clear from the structure that only the A subunit is capable of cleaving the tRNA precursor in this conformation. The catalytic Zn^{2+} ions are absent in the B subunit, and key amino acids are out of range for metal ion coordination (Fig. 4c). His 63 and Asp 211 are rotated by about 90° and 120°, respectively, relative to their positions in the A subunit, and His 140 is displaced by about 8 Å and facing the opposite way. His 247, which may be involved in the positioning of the phosphate ion, is in the region of helix $\alpha 7$ that could not be resolved in the B subunit. There are no significant crystal contacts in the vicinity of the B subunit active site (Supplementary Fig. 3), suggesting that the active site asymmetry is a true conformational state of the dimer in solution, rather than an artefact due to crystal

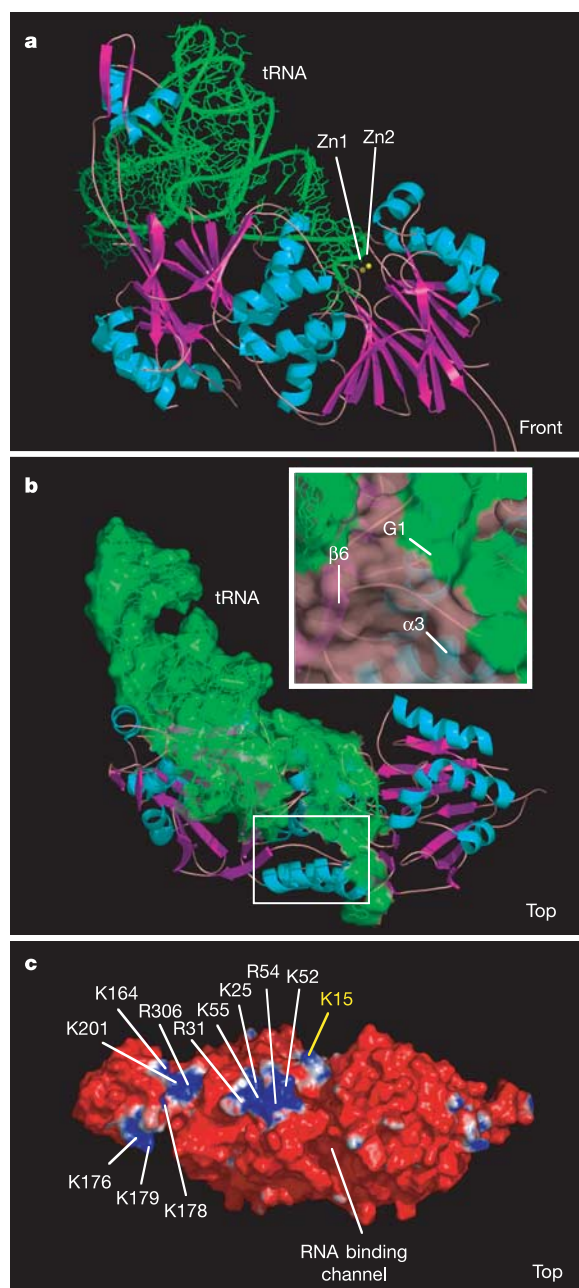


Figure 3 Model of RNase Z-tRNA complex and amino acids important for interaction. **a**, Front view of the RNase Z-tRNA complex. tRNA^{Phe} is shown in green, Zn^{2+} ions as yellow spheres. **b**, Top view. The tRNA is shown as a semi-transparent surface plot. The inset shows a close-up of the boxed region, highlighting the narrowing of the RNA binding channel. The first guanosine residue (G1) of the tRNA and the $\beta 6$ to $\alpha 3$ loop are visible. **c**, Electrostatic potential of RNase Z, with positively charged (+1.0) amino acids in blue and negatively charged (-1.0) amino acids in red. Features of the B and A subunits are labelled in white and yellow, respectively. The RNA binding channel is indicated.

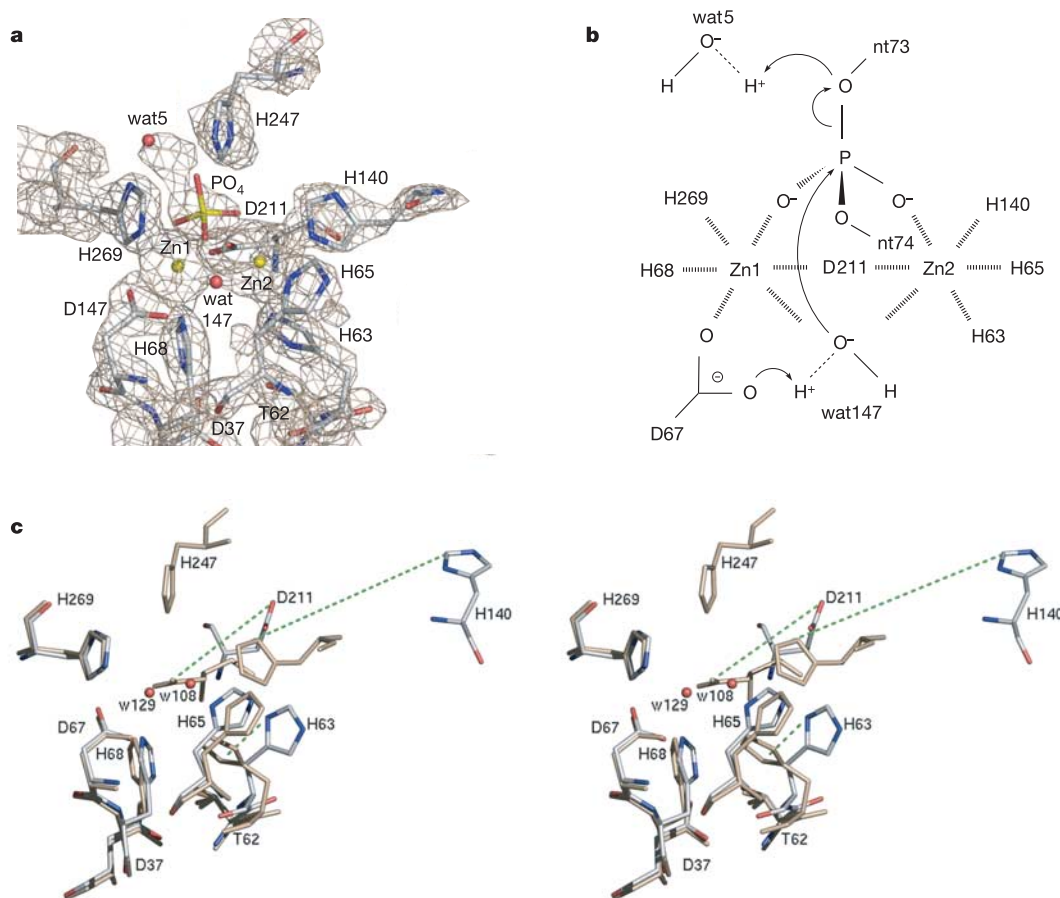


Figure 4 Active sites of RNase Z subunits and the proposed cleavage mechanism. **a**, Active site of A subunit with simulated omit map at 2σ above the mean, in which all labelled features were removed. Wat indicates water molecules. **b**, Proposed reaction mechanism. For description, see text. **c**, Stereo view of superimposed active sites of A and

B subunits. B-subunit features are shown in red and blue, those of the A subunit in beige. Displacement of key residues is indicated by green dotted lines. Water molecules (w) are shown as red spheres. Superposition of active site residues present in both the A and B subunits gives an r.m.s.d. of 4.73 Å.

packing. Although it is possible that a dimer of RNase Z cleaves only a single tRNA precursor molecule at a time, a model whereby the binding of the tRNA results in a conformational change that activates the catalytic site of the B subunit would better account for the allosteric properties of the enzyme. Activation of the B subunit catalytic site would allow coordination of Zn^{2+} and the potential to cleave a second tRNA. Confirmation of this model awaits crystallography studies of the RNase Z–precursor tRNA complex.

RNase Z exists in two forms in nature—a short form, and a longer form found only in eukaryotes, with some species having both enzyme types. It has been suggested that the longer form resulted from a duplication of the smaller enzyme¹⁰. Given the dimeric nature of the smaller protein, it is tempting to speculate that the longer version will prove to be a monomer that resembles the overall protein structure solved here. The N-terminal half of the long form lacks the network of histidine and aspartate residues required for Zn coordination, and can presumably only cleave one tRNA at a time. It is worth noting, in this regard, that the long form of RNase Z behaves as a Michaelian rather than an allosteric enzyme¹².

The β -lactamase domain has proven to be very versatile in terms of the types of covalent bonds it can split in nature. At least 16 families of proteins have exploited the β -lactamase domain for hydrolysis reactions¹⁸. In the class B β -lactamases, it cleaves the C₈–N₅ bond of β -lactam antibiotics¹⁹. In glyoxalase II, it hydrolyses the thioester of S-D-lactoglutathione¹⁵. The β -lactamase domain is also found linked to a flavodoxin-like domain in

rubredoxin:oxygen oxidoreductase (ROO), where it functions as a dehydrase, reducing oxygen to water²⁰. In the case of RNase Z presented here, a flexible arm has been added to the core β -lactamase domain to permit tRNA recognition and hydrolysis of a phosphodiester bond. It is clear that this architecture is suited to many biological hydrolysis reactions once the appropriate substrate recognition domain has been added. □

Methods

Crystallization and data collection

Preliminary crystallization trials were performed at 293 K by hanging-drop vapour diffusion, using a TECAN pipetting robot and Jena Bioscience crystal screens (JBScreen HT I and HT II). Crystals of seleno-methionine-labelled RNase Z (Supplementary Methods 1) were obtained in 30% PEG MME 2K, 0.2 M ammonium sulphate and 0.1 M Na-MES pH 6, using a 2:1 mixture of the protein (8 mg ml⁻¹) and reservoir solution. Crystals of a maximum size of 200 μm appeared within 2 weeks.

X-ray data were collected from cryo-cooled crystals using 25% glycerol as cryoprotectant. MAD data were collected using a MarCCD detector at beam-line BM30A of the European Synchrotron Radiation Facility, Grenoble. RNase Z crystals belong to space group C2221. Assuming two molecules per asymmetric unit, the volume of the asymmetric unit divided by the protein's molecular weight (Matthews' V_m coefficient) is 2.6 Å³ Da⁻¹ and the solvent content is 52% (refs 21, 22). Diffraction data were processed using the DENZO and SCALEPACK programs²³.

Structure determination and refinement

The positions of the Se atoms were located using the SHELXD programme^{24,25} and the automated procedure to locate the anomalous scatters from MAD data at the optimal resolution of 2.5 Å. The best solution proposed by SHELXD was 11 sites, with correlation coefficient (CC) and CC(weak) values of 34.9% and 25.6%, respectively. This solution and its enantiomorph were processed by SHELXE to estimate the phases and the corresponding weights (figures of merit, FOM). The values for contrast, connectivity and

pseudo-free CC of the unambiguous solution were 67%, 93% and 70.0%, respectively, compared to 38%, 84% and 54.6% for the enantiomorph. The unambiguous solution was then analysed to decide which of the 11 sites corresponded to the 10 Se sites of RNase Z. Non-crystallographic symmetry (NCS) operations from 11 heavy atom sites were found automatically using FINDNCS^{22,26}. This results in one NCS for a maximum of 8 matching pairs. The 8 Se positions were refined and phasing was performed with CNS¹³. The resulting phases were extended to 2.2 Å using the CNS density modification procedure. These phases were used to begin automatic model building and refinement using warpNtrace²⁷ and the peak data set with a resolution of 2.1 Å. The resulting ARP model contained 69.4% of the dimer's backbone atoms and 47.5% of the RNase Z sequence. The R_{factor} and R_{free} values for the refinement at this step were 27.1% and 34.1%, respectively.

The initial model provided by ARP was completed and adjusted with the program O²⁸, first with the solvent-flattened MAD map, then with the combined model and experimental phases (as calculated by CNS). Refinement was performed using CNS and the peak data set at 2.1 Å resolution. Standard protocols of structure refinement with experimental phases (Hendrickson–Lattman coefficients from MAD phasing), including energy minimization, both before and after simulated annealing (using torsion angle dynamics), bulk solvent correction and anisotropic B -factors (applies the isotropic component of the correction to the model and the anisotropic component to the observed data) were used. Each round of refinement was alternated with a round of manual refitting using the cross-validated, sigma- A weighted, phase combined $2F_o - F_c$ and $F_o - F_c$ maps. Progress in the model refinement was evaluated by decrease in the free R -factor.

The current model includes 566 residues (1–158, 197–307 for the A subunit, and 1–236, 247–307 for the B subunit), two Zn^{2+} ions, a phosphate ion and 176 water molecules. Residues 159–196 of the A subunit and 237–246 of the B subunit were not visible in the electronic density maps, and are omitted in the final molecular model. Superposition of 518 α -carbon atoms from the A and B subunits (residues 1–158, 197–236 and 247–307) results in an r.m.s.d. of 1.76 Å.

Received 24 September; accepted 16 December 2004; doi:10.1038/nature03284.

Published online 16 January 2005.

1. Frank, D. N. & Pace, N. R. Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annu. Rev. Biochem.* **67**, 153–180 (1998).
2. Garber, R. L. & Altman, S. *In vitro* processing of *B. mori* transfer RNA precursor molecules. *Cell* **17**, 389–397 (1979).
3. Li, Z. & Deutscher, M. P. Maturation pathways for *E. coli* tRNA precursors: a random multienzyme process *in vivo*. *Cell* **86**, 503–512 (1996).
4. Garber, R. L. & Gage, L. P. Transcription of a cloned *Bombyx mori* tRNA2Ala gene: nucleotide sequence of the tRNA precursor and its processing *in vitro*. *Cell* **18**, 817–828 (1979).
5. Schiffer, S., Rosch, S. & Marchfelder, A. Assigning a function to a conserved group of proteins: the tRNA 3'-processing enzymes. *EMBO J.* **21**, 2769–2777 (2002).
6. Nashimoto, M. Distribution of both lengths and 5' terminal nucleotides of mammalian pre-tRNA 3' trailers reflects properties of 3' processing endoribonuclease. *Nucleic Acids Res.* **25**, 1148–1154 (1997).
7. Kunzmann, A., Brennicke, A. & Marchfelder, A. 5' end maturation and RNA editing have to precede tRNA 3' processing in plant mitochondria. *Proc. Natl Acad. Sci. USA* **95**, 108–113 (1998).
8. Schierling, K., Rosch, S., Rupprecht, R., Schiffer, S. & Marchfelder, A. tRNA 3' end maturation in archaea has eukaryotic features: the RNase Z from *Haloferax volcanii*. *J. Mol. Biol.* **316**, 895–902 (2002).
9. Pellegrini, O., Nezzar, J., Marchfelder, A., Putzer, H. & Condon, C. Endonucleolytic processing of CCA-less tRNA precursors by RNase Z in *Bacillus subtilis*. *EMBO J.* **22**, 4534–4543 (2003).
10. Tavtigian, S. V. *et al.* A candidate prostate cancer susceptibility gene at chromosome 17p. *Nature Genet.* **27**, 172–180 (2001).
11. Minagawa, A., Takaku, H., Takagi, M. & Nashimoto, M. A novel endonucleolytic mechanism to generate the CCA 3' termini of tRNA molecules in *Thermotoga maritima*. *J. Biol. Chem.* **279**, 15688–15697 (2004).
12. Mohan, A., Whyte, S., Wang, X., Nashimoto, M. & Levinger, L. The 3' end CCA of mature tRNA is an antideterminant for eukaryotic 3'-tRNAse. *RNA* **5**, 245–256 (1999).
13. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
14. Schiffer, S., Helm, M., Theobald-Dietrich, A., Giege, R. & Marchfelder, A. The plant tRNA 3' processing enzyme has a broad substrate spectrum. *Biochemistry* **40**, 8264–8272 (2001).
15. Cameron, A. D., Ridderstrom, M., Olin, B. & Mannervik, B. Crystal structure of human glyoxalase II and its complex with a glutathione thiolester substrate analogue. *Struct. Fold. Des.* **7**, 1067–1078 (1999).
16. Nashimoto, M., Wesemann, D. R., Geary, S., Tamura, M. & Kaspar, R. L. Long 5' leaders inhibit removal of a 3' trailer from a precursor tRNA by mammalian tRNA 3' processing endoribonuclease. *Nucleic Acids Res.* **27**, 2770–2776 (1999).

17. Vogel, A., Schilling, O., Niecke, M., Bettmer, J. & Meyer-Klaucke, W. *ElaC* encodes a novel binuclear zinc phosphodiesterase. *J. Biol. Chem.* **277**, 29078–29085 (2002).
18. Daiyasu, H., Osaka, K., Ishino, Y. & Toh, H. Expansion of the zinc metallo-hydrolase family of the β -lactamase fold. *FEBS Lett.* **503**, 1–6 (2001).
19. Carfi, A. *et al.* The 3-D structure of a zinc metallo-beta-lactamase from *Bacillus cereus* reveals a new type of protein fold. *EMBO J.* **14**, 4914–4921 (1995).
20. Frazao, C. *et al.* Structure of a dioxygen reduction enzyme from *Desulfovibrio gigas*. *Nature Struct. Biol.* **7**, 1041–1045 (2000).
21. Matthews, B. W. Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497 (1968).
22. CCP4. The CCP4 Suite: Programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
23. Otwinowski, Z. & Minor, W. in *Methods in Enzymology* Vol. 276, *Macromolecular Crystallography Part A* (eds Carter, C. W. Jr & Sweet, R. M.) 307–326 (Academic, New York, 1997).
24. Uson, I. & Sheldrick, G. M. Advances in direct methods for protein crystallography. *Curr. Opin. Struct. Biol.* **9**, 643–648 (1999).
25. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D* **58**, 1772–1779 (2002).
26. Lu, G. FINDNCS: A program to detect non-crystallographic symmetries in protein crystals from heavy atom sites. *J. Appl. Crystallogr.* **32**, 365–368 (1999).
27. Perrakis, A., Morris, R. & Lamzin, V. S. Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6**, 458–463 (1999).
28. Kleywegt, G. J. & Jones, T. A. in *From First Map to Final Model* (eds Bailey, S. R. H. & Waller, D.) 59–66 (SERC Daresbury Laboratory, Daresbury, UK, 1994).
29. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Springer, D. Picot, R. Giégé, F. Allemand, V. Arluison and F. Dardel for discussions, S. Fieulaine and M. Pirocchi for help with the beam-line BM30A at the European Synchrotron Radiation Facility, and J.L. Popot for use of crystallography facilities and X-ray generator at the IBPC. This work was supported by the CNRS (UPR 9073), Université Paris VII-Denis Diderot, PRFMMP 2001/2003, and ACI Jeunes Chercheurs from the Ministère de l'Éducation Nationale.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.C. (condon@ibpc.fr). Coordinates are deposited in the Brookhaven Protein Data Bank under accession code 1Y44.

corrigendum

The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II

Minkyu Kim, Nevan J. Krogan, Lidia Vasiljeva, Oliver J. Rando, Eduard Nedea, Jack F. Greenblatt & Stephen Buratowski

Nature **432**, 517–522 (2004).

In this Letter to *Nature*, reference 5 should be: Connelly, S. & Manley, J. L. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev.* **2**, 440–452 (1988). □

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Andy Douglas (4975)
Nils Moeller (4953)
Irene Vigila-Atton (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Advertising Production

Manager: Billie Franklin
To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:
Tom Hancock

Naturejobs online production:
Niamh Shields

European Satellite Office

**Germany/ Austria/ Italy/
The Netherlands/ Belgium:**

Patrick Phelan
Tel +49 89 54 90 57 11
Fax +49 89 54 90 57 20
e-mail: p.phelan@nature.com
Sharon de Weert
Tel +44 (0) 20 7843 4970
e-mail: s.deveert@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harakakamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746
Asia-Pacific Sales Director:
Rinoko Asami
e-mail: rasami@naturejpn.com

naturejobs

Reversal of fortune

A new rule that severely limits employees of the US National Institutes of Health (NIH) from pursuing outside income could have a cooling effect on recruitment to its 18,000-member workforce. It could also have repercussions farther afield, as institutions choose either to adopt a similar policy or to exploit it to recruit scientists who might have headed for the NIH.

Rules about accepting outside income, including pharmaceutical and biotech stock for consulting work, were relaxed in 1995 under the NIH's then-director Harold Varmus, in part to increase recruitment of high-level scientists. The revision came under fire last year thanks to a few dozen instances in which scientists apparently didn't disclose potential conflicts of interest. Both Varmus and current NIH director Elias Zerhouni said last year that only senior staff, such as individual NIH institute directors, should be banned from receiving outside income.

But the latest rule change, announced last week, marks another reversal, because it extends the ban to all employees. Zerhouni has said this may have some effect on recruiting top talent — the institute's campus in Bethesda, Maryland, is the biggest employer of life scientists in the greater Washington DC area (see page 664). And the move seems counterintuitive as Zerhouni's 'roadmap' for the NIH calls for an increase in collaborations between academia and industry.

Some NIH employees have said that enforcing the existing system would have been fairer than exercising strict limits on thousands of scientists who have played by the rules. In the past decade or so, it has become common for scientists to live professionally in several worlds at once. This arrangement creates better career paths for scientists. But for the immediate future, for scientists wanting to work at the interface between the business and academic worlds, such a path won't lead to Bethesda.

Paul Smaglik
Naturejobs editor



Contents

REGIONS

Building for success
in Washington DC p664

CAREER VIEW

Bricks & Mortar
Scottish Structural
Proteomics Facility
Graduate Journal
A rewarding journey
Movers
Peter Calow p666

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Capital collaboration

Washington DC

Bioomedical scientists across the United States feel the effects when the National Institutes of Health (NIH) has its budget cut or increased. Like others in the field, those at the NIH's main campus in Bethesda, Maryland, and its neighbours in the Washington DC region, are bracing themselves for flat budgets. Yet interdisciplinary opportunities continue to grow in the area.

The NIH's Porter Neuroscience Research Center is a prime example. Opened last June, the centre brings together researchers from several NIH institutes on the basis of thematic research interests such as synapses, cognition and mood — the first interdisciplinary arrangement of its kind on the NIH campus.

Cutting-edge equipment and a solid start-up package attracted John Isaac, a recently arrived senior investigator who studies the molecular mechanisms of developmental synaptic plasticity. "The NIH upfront could offer me an entire budget to run my lab for five years at least," says Isaac, who came from the University of Bristol, UK. Typically, neither UK nor US universities can match such a commitment.

But the institute has so far made few appointments from outside the NIH campus. Only the first phase of the centre is complete; a second phase is planned, and would offer more jobs, but it is not yet fully funded. It's unclear when funding for the next phase will appear, says Story Landis, director of the National



Norma Allewell (top) and Gerald Rubin are seeking scientists with interdisciplinary skills.

Institute of Neurological Disorders and Stroke (NINDS).

Other NIH institutes are showing targeted growth, or at least some employment opportunities. Flush with money for biodefence research, the NIH campus will also have a new biodefence research facility, along with similar labs at satellites in Frederick, Maryland, and Rocky Mountain, Montana. Because the biodefence labs will have biosafety level two and three capability (biosafety level four in the case of the Rocky Mountain labs), the National Institute of Allergy and Infectious Diseases (NIAID) is looking for people with training in biocontainment, says Kathy Zoon, deputy director of planning and development with the NIAID's intramural programme. Although the NIH isn't likely to be taking on a lot more people, routine turnover and retirements do allow for some openings. Landis notes that the National Institute of Mental Health and NINDS are looking for new directors. Several institutes, including

NINDS, have openings for scientists interested in overseeing the grant-review process.

And even in NIH institutes not due for growth, the Bethesda campus remains a great place for young scientists seeking training. The campus hosts 1,000 summer students, 300 graduate students through collaborative agreements with universities, 100 medical students, 500 post-baccalaureate students, and 2,800 postdoctoral fellows. Almost all NIH-trained scientists find work in academic and industry positions, according to Michael Gottesman, deputy director for intramural research at the NIH. He notes that the number of newly hired tenure-track investigators has dropped from 49 in 2000 to 34 in 2004, reflecting the flattening of the NIH budget following its doubling in the early 2000s.

Despite struggling with its own budget constraints, the University of Maryland has expanded its biology and chemistry research efforts in recent years. A \$62-million bioscience research building, due for completion in 2006, will house 35 new tenure-track biology faculty members. The college is seeking biologists with interdisciplinary interests, says Norma Allewell, dean of the University of Maryland College of Life Sciences. An additional \$23-million chemistry wing will bring in ten new chemists and biochemists in the next few years. The university is also helping to establish a technology park, which will include biotech ventures, a mile from the campus.

Like many states struggling to balance budgets, Maryland has reduced funding for its public

An emerging market

In 2000, business speculation on biotechnology was centred on the suburbs of Washington DC. Companies such as Celera seemed poised to exploit the bounty of the nearly sequenced human genome. The firms were anticipating huge profits, and rapid employment growth had made the region the third-largest biotech market, after the San Francisco Bay area and Boston/Cambridge.

But the bubble burst in 2001, and the valuation and employment capacity of these companies have failed to live up to expectations. "The trend now is towards biopharmaceutical companies — those actually developing drugs," says Rene Salas, who oversees the mid-Atlantic region for analysts Ernst & Young. But the bubble hasn't burst completely. Although the mid-Atlantic region slipped from third to fourth nationally, the region's biotech sector has still experienced growth.

The upshot is that Maryland, Virginia and Washington DC still constitute an emerging market for biotechnology companies, says Salas. Only a few companies have reported revenue, even fewer have reported actual profit, and none went public in 2004. "The past year was a pretty quiet year for Maryland," says Salas. But he adds that a few companies in the area look as though they could go public, and that the financing totals for 2004 were second only to those for 2000.

E.R.



Work in progress: the new chemistry wing at the University of Maryland (top) and the Howard Hughes Medical Institute's Janelia Farm campus (left) are bringing fresh opportunities to the DC area.

universities. Long-term funding cuts would result in the department offering fewer faculty positions rather than making start-up packages for recruits smaller, says Allewell. Sarah Tishkoff, an assistant professor of biology at the University of Maryland, finds budget cuts disheartening, although she and her department have not been as affected as others. "I think what it did more than anything is hurt morale," says Tishkoff.

WIDE FOCUS

Thirty miles to the north, Johns Hopkins University Medical School in Baltimore is currently the top recipient of NIH extramural funds and continues to expand. Again, interdisciplinary research is the focus. Hopkins is expanding its new Institute for Basic Biomedical Sciences, adding four yet-to-be-determined programme areas, each with four or five investigators. The institute's additions will be part of a planned life-sciences technology park in Baltimore that will house both Hopkins investigators and researchers from companies. The first of two life-science park buildings will be completed in 2007.

Thirty miles west of DC, the Howard Hughes Medical Institute (HHMI) in Chevy Chase, Maryland, is also launching a major interdisciplinary effort. Slated to open in the summer of 2006 with 12–15

group leaders, the HHMI Janelia Farm Research Campus in Loudon County, Virginia, will eventually house 24 group leaders in an open-lab setting. As with current HHMI investigators, Janelia researchers will be fully funded for five years, then reviewed for renewal. By 2010, Janelia is expected to have 24 group leaders and a total research staff of 200 to 300.

Unlike the University of Maryland, the new facility will not simply be seeking biologists with interests in other fields, says director Gerald Rubin. Instead, it will put physicists, engineers and computer scientist PhDs with an interest in life sciences alongside biologists in the hope of establishing interdisciplinary collaborations on high-risk research. "Our basic attitude is: 'If you're willing to bet a career we're willing to bet a few million dollars,'" says Rubin.

Of course, prospective researchers have very different work environments to choose from. Those at Janelia, like NIH intramural researchers, are free from teaching duties and constant grant writing. On the other hand, they're not part of a diverse university community complete with humanities departments and sports teams — an attraction for some. Some may not embrace the modern, hypercollaborative open-lab setting of places such as the Porter centre, Janelia and some of the new Hopkins labs. And NIH researchers are severely restricted in terms of collaborations with industry. Landis notes, though, that the biggest obstacle for many recruits is the high price of housing in the DC area.

DC area institutions such as the University of Maryland and Hopkins hope to expand industry collaborations as biotech in the area continues to grow (see 'An emerging market', opposite). With the new interdisciplinary centres and a rise in the number of biotech companies, the DC area should see an overall growth in the number of scientific jobs this year.

Eugene Russo is a science writer based in Takoma Park, Maryland.

GRADUATE JOURNAL

A rewarding journey

Accepting a studentship at the International Institute of Molecular and Cell Biology in Warsaw, Poland, was a crucial step for my training. But in order to take advantage of it, I've entered what I call my 'commuting matrix'. As a student without any real income, I can't afford to rent a flat in the capital, so I have chosen to live in my home town of Lodz and travel to and from Warsaw every weekday to save money.

The 120-km train trip takes two hours each way. To make the most of this time, I read scientific articles on the train. And even though I work in the high-tech field of bioinformatics, I use low-tech pen and paper to scribble down scientific problems, and to plan my day and work out longer-term goals.

When I get to Warsaw, it all seems worthwhile. Working with a leading team of researchers is not only a test but also great fun. There is a friendly atmosphere and I am surrounded by helpful people and have a boss who believes in giving young scientists a chance.

I wish that more students could work with such helpful teams and supervisors, who see that the future lies in the hands of young people like me — and who are prepared to make an effort to break down the barriers that stand in our way. ■

Karolina Tkaczuk is a graduate student at the Technical University of Lodz, Poland.

BRICKS & MORTAR

Scottish Structural Proteomics Facility, University of St Andrews, UK

A £6-million (US\$11-million) lab dedicated to finding treatments for 'superbug' infections — those resistant to conventional antibiotics — opened at the University of St Andrews in Scotland in December.

The Scottish Structural Proteomics Facility (SSPF) is a collaboration between the universities of Dundee and St Andrews. It is designed to streamline the process of drug design, from the identification of novel therapeutic targets for drug-resistant bacteria to producing candidate drug leads.

The centre emerged from the realization that Britain had fallen behind Japan and the United States in this research field, says its co-director James Naismith, professor of chemical biology at the



University of St Andrews. It is intended to close a gap in both skills and technology.

Essentially a structural-proteomics facility dedicated to drug discovery, the SSPF will house industrial-scale equipment for crystallography and nuclear magnetic resonance spectroscopy, with an emphasis on expressing difficult-to-produce membrane proteins. It will also have large-scale robotic equipment with cloning and expression technology.

Naismith says that this technology will make it easier for him to recruit international talent. Already there are signs that the SSPF is helping to close the skills gap, as it has attracted scientists from Japan and the United States. "We are now recruiting for other principal investigators,"

Naismith says.

Researchers will collect the findings on a database of enzymes to share with colleagues around Britain. They are especially interested in viral entry to cells and replication.

The centre was initiated with a development grant from the Scottish Higher Education Funding Council, and established with £600,000 from the two universities and £4.2 million from the Biotechnology and Biological Sciences Research Council. ■

Paul Smaglik is editor of Naturejobs.

MOVERS

Peter Calow, director, Environmental Assessment Institute, Copenhagen, Denmark



The career of Peter Calow has undergone two large changes of direction. In 1984 the British zoologist left basic research — his PhD from the University of Leeds was on the population ecology of aquatic snails — to work on applied ecotoxicology and environmental risk management.

Twenty years on, Calow has made another big move, this time to environmental policy assessment. In November, the 57-year-old professor at the University of Sheffield was appointed the new director of the Danish

Environmental Assessment Institute (EAI) in Copenhagen. For the next five years, he will be responsible for socioeconomic and cost-benefit analyses of a wide range of environmental policy options.

"My career has been a natural progression from experimental science to the policy arena," he says. "I think that this is now the most exciting job I have ever had."

The EAI was set up by the Danish government just two years ago, as an independent environmental think-tank. Thanks to its prominent founding director, Bjørn Lomborg, the small institute has quickly received attention from far beyond Denmark.

Calow appreciates the contribution of his predecessor. Lomborg, he says, asked important new questions — a practice that Calow wants to continue. He knows all about the controversy associated with using economic approaches to environmental problems.

But thorough cost-benefit analysis is crucial to finding the best policy solutions to many environmental issues, from chemical legislation to climate change, he says.

As a PhD student, Calow was greatly influenced by the work of Nobel laureate Peter Medawar, particularly his book *The Art of the Soluble*. "Medawar taught me a lot about how to ask the right questions and design experimental approaches to answer them." No wonder, then, that Calow has always advised his students to read the book.

Defining the right problems and questions will keep Calow busy for a while, as he and his staff develop a strategic plan for the EAI. Although priorities may shift, the institute's core activities will remain, he says. "We want to influence policy-makers and the public, and give inspiration to the academic community in Denmark, Europe and the whole world." ■

CV

1989–2004: Professor of zoology, University of Sheffield, UK.

1996–2003: Director, Environmental Businesses Network, Sheffield, UK.

1991–95: Director of the Institute of Environmental Sciences and Technology, Sheffield, UK.

1984–1989: Head of the zoology department, University of Sheffield, UK.

Under martian ice

Cold ... and never more alone.

Stephen Baxter

I suppose it was the Fermi paradox that drew me into science in the first place. I never expected to find a resolution to the paradox — not here! But that, it seems, is what lies beneath the ancient ice of Mars.

It's a shame that this discovery is hogging all the headlines back home, for our assault on the martian South Pole is an epic in its own right. The ice here is old and dirty, polluted with traces that might let us reconstruct Mars's climatic history, just as on Earth. And the morphology is extraordinary, with vast canyons spiralling out from a central dome, a self-organizing system 1,000 kilometres across.

But the adventure here is very human. We're deep into our first Earth-year-long polar winter. We're occupying ourselves with sample analysis, mailing home, teaching one another languages — would you believe a dozen spoken fluently by our polyglot six-strong crew? In other words we're over-wintering, just like Shackleton.

And, along with the rest of humanity, we're mulling over the implications of our discovery under the ice.

The bedrock under the ice is among the oldest on Mars. One of our key objectives here was to assemble detailed maps of the hidden landscape with sounding radar and active seismology. What we found was something utterly unexpected.

It's a quite inhuman layout, of low walls outlining pentagonal and hexagonal areas, within which boxy structures huddle. We don't know its purpose, but the 'city' under the ice is unmistakable as a sign of intelligence. Somebody has been here before us.

Of course these structures are not native. Eighty years after the first successful robot landers, we were sure that life here never advanced beyond stromatolites. I have a sneaking feeling that whoever built the city wasn't so terribly unlike us. They were drawn by the adventure of reaching the pole, just as we were.

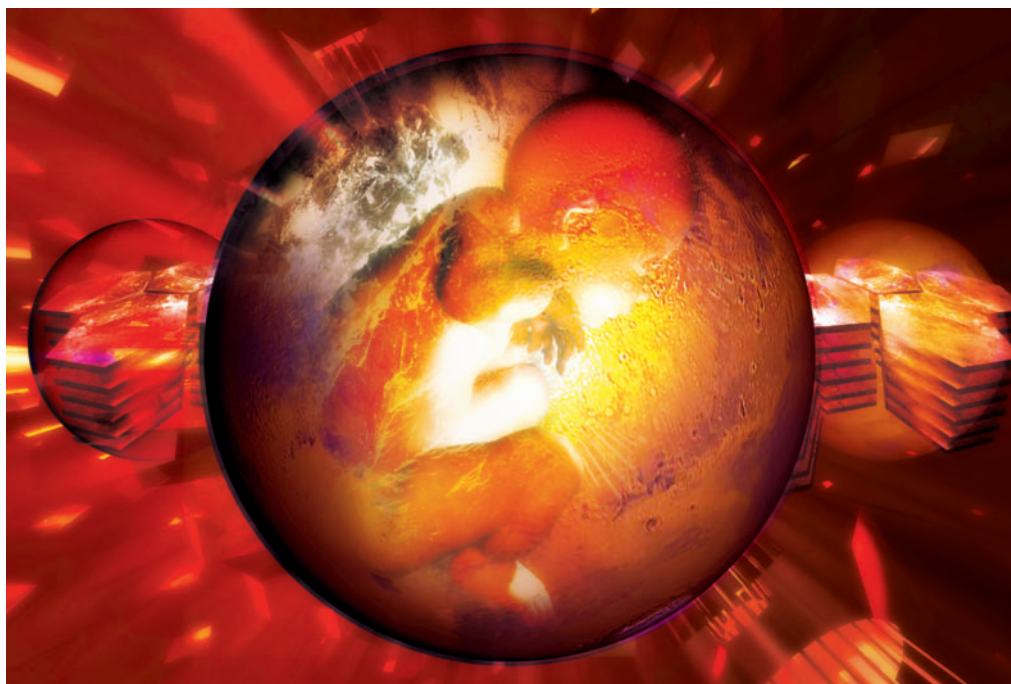
But it drives me crazy that we'll probably never know. It's already clear that their visit must have been long ago — long even for a geologist like me, billions of years back. Which is where Fermi comes in.

I'm sure you're familiar with the question Enrico Fermi asked in 1950: "Where is everybody?" If extraterrestrial aliens exist, they should have spread everywhere by now. So how come we don't see them? Our vision of the Universe has expanded greatly since 1950, but we've still turned up no incontrovertible evidence of intelligence away from Earth. Until now.

In retrospect, we should have expected to find traces of long-gone travellers. Interstellar visits were actually more likely in ancient times than now. The Galaxy's peak star-formation rate seems to have been some

did come long ago, they didn't visit a roiling young Earth, but the relatively advanced biosphere of Mars.

We talk of nothing else. Our key activity next summer was to have been our ice core, kilometres long, all the way down to the basement rock. Now we don't care about Mars; we're diverting the core to try to get a sample of the 'city'. The biologists are excitedly debating how to distinguish any traces of life. But if the alien visitors practised Planetary Protection Protocols as scrupulously as we do, they may have left no trace of themselves at all.



five billion years ago — just before the birth of the Sun — so most stars and planetary systems must be older than our own. The Galaxy's climax as an arena for nurturing civilization was deep in the past.

And if they did come to the Solar System so long ago, where would they have visited?

Early Mars was more hospitable to life than Earth. Being smaller, Mars cooled quicker, and life made an earlier start. Mars was less of a target for the planet-sterilizing impactors that roamed the young Solar System. Young Mars even enjoyed an atmosphere rich in oxygen. Indeed, as everybody knows by now, we've confirmed that the original source of life on Earth was in fact Mars, transmitted by impact-detached meteorites.

There's the resolution to Fermi, at last. We don't see anybody because they have long gone, their worlds exhausted. And when they

It's driving me crazy. I'm bending mission rules to do it, but I go out in sleep periods, and walk away from the base, and think.

We humans are just not used to being alone. We evolved in a world full of non-human hominids, other kinds of mind. That's why we fill the sky with demons and aliens; we can't stand the echoing silence we have created. And now we know that we will find nothing out there among the stars but exhausted worlds, and museums, and ruins.

I keep these thoughts to myself. The martian winter is long, and morale is everything. I let it all out on the frozen air, under Jupiter's wheeling moons, a 50-year-old geologist as mixed up as a 10-year-old. Then I walk back to the human warmth of our base.

Stephen Baxter's most recent novel is *Exultant* (Gollancz, 2004). He lives in Northumberland, UK.